2017

Defining Sites Of Replication Fork Collapse Caused By Atr Inhibition

Nishita Kalpendu Shastri
University of Pennsylvania, nishita@mail.med.upenn.edu

Defining Sites Of Replication Fork Collapse Caused By Atr Inhibition

Abstract
DEFINING SITES OF REPLICATION FORK COLLAPSE CAUSED BY ATR INHIBITION

Nishita K. Shastri

Eric J. Brown

Replication stress, characterized by stalling of DNA replication and the accumulation of abnormal replication intermediates, has been linked to the genomic instability observed in cancer. Previous studies have defined specific genomic sequences that are difficult to replicate to be more vulnerable to replication-associated breaks and rearrangements. However, many of these sequences have been identified through indirect and potentially biased approaches. To identify DNA sequences that contribute to replication-associated genomic instability, I will describe genome-wide screens I have performed to determine the location, sequence, and frequency of replication perturbations within the mammalian genome upon replication stress. Ataxia telangiectasia and Rad3-related protein (ATR) is a checkpoint kinase that is a key upstream regulator of the response pathway to replication fork stalling during replication stress that prevents fork collapse.

Through inhibition of this response pathway in mouse embryonic fibroblasts, my aims are to 1) characterize regions that lead to frequent replication fork stalling and collapse, and 2) further define genomic regions that become processed into double-strand breaks. Since replication protein A (RPA) binds to single-stranded DNA that becomes exposed when replication forks stall, RPA ChIP-Seq has been performed to map sites of frequently collapsed replication forks; however, not all stalled replication forks result in breaks. To differentiate a replication fork that has simply stalled from a fork that has become sensitized to double-strand break formation, I developed and applied a novel and specific break-detection assay, BrITL. With these complementary approaches to map replication-problematic loci, subsequent bioinformatics methods have been utilized to characterize features of the identified genomic regions that make it prone to fork collapse and detrimental DNA break formation when cells experience replication stress.

While well-established difficult-to-replicate sequences (e.g. triplet and telomere repeats) exhibited enhanced fork collapse in RPA ChIP'd cells exposed to replication stress, these sequences were overshadowed by sites composed of previously uncharacterized simple tandem repeats. Circular dichroism and thermal difference absorption spectra indicate that the most commonly observed simple repeat at RPA-enriched sites (CAGAGG) folds into a stable intramolecular secondary structure and is sufficient to stall DNA replication in vitro and in vivo. BrITL analysis confirmed that these repetitive regions of RPA accumulation are also sites of DNA breakage. Interestingly, a majority of break sites identified by BrITL do not associate with RPA accumulation, but rather tend to locate around inverted retroelements that are predicted to form highly stable intrastrand stem-loop structures. Due to the lack of available ssDNA at these potential hairpin-forming sites, RPA accumulation would be limited. Overall, my studies represent the first unbiased identification of mammalian genomic sites that are vulnerable to replication stress and rely on ATR for stability.

Degree Type
Dissertation

DEFINING SITES OF REPLICATION FORK COLLAPSE CAUSED BY ATR INHIBITION

Nishita K. Shastri

A DISSERTATION

in

Pharmacology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

Eric J. Brown, PhD

Associate Professor of Cancer Biology

Graduate Group Chairperson

_____

Julie A. Blendy, PhD, Professor of Pharmacology

Dissertation Committee

Klaus Kaestner, PhD, Thomas and Evelyn Suor Butterworth Professor in Genetics (Chair)

Gerd Blobel, MD, PhD, Frank E. Weise III Professor of Pediatrics

F. Bradley Johnson, MD, PhD, Associate Professor of Pathology and Laboratory Medicine

I would like to dedicate this to my family: my brother, Anujit, my sister, Ankita, and my parents, Kalpendu and Bhavna Shastri. I would also like to dedicate this to my grandparents, Biharilal and Jaya Shah, and Ranjit and Anila Shastri.

They have always been my greatest support and my highest inspiration. I hope I can make them proud.

ACKNOWLEDGMENTS

I would like to thank the following people for all their help:

Yu-Chen Tsai, who started the work with optimization and application of the RPA ChIP-Seq protocol on MEFs, and provided the constructs for the structural assays as well as for the *in vivo* and *in vitro* studies on fork pausing.

Dillon Maloney, who provided the essential bioinformatics work that led to our customized pipeline for ChIP-Seq and BrITL deep sequencing data analysis. He was also integral in developing the program REQer for our in-depth look at repetitive sequences in our ChIP-Seq data sets.

Our bioinformatics collaborators (Jonathan Schug, Rafael Casellas, Marei Dose) who provided key analyses on our sequences and advice on demonstrating the significance of our findings.

Our collaborators for their work on demonstrating both the formation of unique structures by our identified sequences (Liliya Yatsunyk, Jessica Chen, Deondre Jordan, Barrett Powell) and the resultant stalling of the polymerase complex in our *in vitro* and *ex vivo* systems (Kristin Eckert, Suzanne Hile).

The Functional Genomics Core (FGC) at Penn for deep sequencing RPA ChIP-Seq and BrITL samples, and the Institute for Biomedical Informatics (IBI) at Penn for work done to identify inverted repeats in the mouse genome.

The Brown lab members for their considerable support and advice, and, most of all, Eric Brown, for his incredible mentorship.

ABSTRACT

DEFINING SITES OF REPLICATION FORK COLLAPSE CAUSED BY ATR INHIBITION

Nishita K. Shastri

Eric J. Brown

Replication stress, characterized by stalling of DNA replication and the accumulation of abnormal replication intermediates, has been linked to the genomic instability observed in cancer. Previous studies have defined specific genomic sequences that are difficult to replicate to be more vulnerable to replication-associated breaks and rearrangements. However, many of these sequences have been identified through indirect and potentially biased approaches. To identify DNA sequences that contribute to replication-associated genomic instability, I will describe genome-wide screens I have performed to determine the location, sequence, and frequency of replication perturbations within the mammalian genome upon replication stress. Ataxia telangiectasia and Rad3-related protein (ATR) is a checkpoint kinase that is a key upstream regulator of the response pathway to replication fork stalling during replication stress that prevents fork collapse.

Through inhibition of this response pathway in mouse embryonic fibroblasts, my aims are to 1) characterize regions that lead to frequent replication fork stalling and collapse, and 2) further define genomic regions that become processed into double-strand breaks. Since replication protein A (RPA) binds to single-stranded DNA that becomes exposed when replication forks stall, RPA ChIP-Seq has been performed to map sites of frequently collapsed replication forks; however, not all stalled replication forks result in breaks. To differentiate a replication fork that has simply stalled from a fork

that has become sensitized to double-strand break formation, I developed and applied a novel and specific break-detection assay, BrITL. With these complementary approaches to map replication-problematic loci, subsequent bioinformatics methods have been utilized to characterize features of the identified genomic regions that make it prone to fork collapse and detrimental DNA break formation when cells experience replication stress.

While well-established difficult-to-replicate sequences (e.g. triplet and telomere repeats) exhibited enhanced fork collapse in RPA ChIP'd cells exposed to replication stress, these sequences were overshadowed by sites composed of previously uncharacterized simple tandem repeats. Circular dichroism and thermal difference absorption spectra indicate that the most commonly observed simple repeat at RPA-enriched sites (CAGAGG) folds into a stable intramolecular secondary structure and is sufficient to stall DNA replication *in vitro* and *in vivo*. BrITL analysis confirmed that these repetitive regions of RPA accumulation are also sites of DNA breakage. Interestingly, a majority of break sites identified by BrITL do not associate with RPA accumulation, but rather tend to locate around inverted retroelements that are predicted to form highly stable intrastrand stem-loop structures. Due to the lack of available ssDNA at these potential hairpin-forming sites, RPA accumulation would be limited. Overall, my studies represent the first unbiased identification of mammalian genomic sites that are vulnerable to replication stress and rely on ATR for stability.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

ABSTRACT iv

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER 1: INTRODUCTION 1

1.1 Replication stress 2

1.2 Replication checkpoint response 7

1.3 Replication fork collapse and recovery 8

1.4 Cancer genomic instability 10

1.5 Break-identification methods 11

1.6 Significance of defining genomic sites that rely on ATR for stability 15

1.7 Aims 17

References 21

Contributions of work presented in Chapter 2 27

CHAPTER 2: RPA CHIP-SEQ ON REPLICATION-STRESSED MEFs 28

2.1 Genome-wide identification of RPA-enriched sites following ATR inhibition 28

2.2 REQer: Enrichment of repetitive sequences in RPLs ……..… 42

2.3 Simple tandem repeats in RPLs form stable intrastrand structures …….. 51

2.4 Simple tandem repeats in RPLs lead to fork stalling *in vitro* and *ex vivo* ……. 61

References ……... 71

CHAPTER 3: BrITL ON REPLICATION-STRESSED MEFs ……..… 74

3.1 Development and validation of BrITL, DNA break-detection assay …….. 74

3.2 Simple tandem repeats in RPLs undergo double-strand breakage ……..… 78

3.3 BrITL-specific sites are composed of hairpin-forming inverted retroelements …... 82

References ……... 93

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS ……..... 95

4.1 RPA ChIP-Seq identification of fork-collapse sites ……..….. 96

4.2 BrITL identification of fork-collapse and break sites ……..…… 98

4.3 Models for fork-collapse ……..……. 100

4.4 Perspectives ……... 102

References ……..…….. 105

APPENDIX ……..…….. 107

Cell lines ……..……..…. 107

Cell treatments ……..…… 107

Cell culture ……..…… 107

RPA ChIP-Seq ……. 107

BrITL …….. 108

Bioinformatics …….. 112

Fork-pausing ……..….. 115

Biophysical characterization of DNA secondary structures ……..…. 118

References ……..…….. 121

LIST OF TABLES

Table 1. Genomic features of RPLs ……..… 36

Table 2. Comprehensive list of RPL peaks and associated simple repeats that overlap with CTCF binding sites …….. 37

Table 3. Summary of repeats found enriched in the ATRi+aph^{18hrs} RPA ChIP ……... 41

Table 4. Biophysical parameters of RPL repeat sequences …….. 54

Table 5. UV-vis melting data on increasing lengths of CAGAGG repeat ……. 59

Table 6. Inverted repeats in mouse genome ……..... 87

LIST OF FIGURES

Figure 1. Hotspots for replication-induced rearrangements ……..... 2

Figure 2. Schematic of replication stress-induced fork stalling ……..... 29

Figure 3. Schematic of RPA ChIP-Seq method ……... 30

Figure 4. RPA ChIP-Seq coverage and ratio tracks on different chromosomes ……... 32

Figure 5. Venn diagram overlap of peaks identified from each ATRi condition ……. 34

Figure 6. Tiered categorization of RPL peaks …….. 35

Figure 7. Overlap of RPL peaks with CTCF binding sites …….. 37

Figure 8. RPA peaks along part of chromosome 6 of the mouse genome ……. 41

Figure 9. Complex repeat enrichment in RPA ChIP samples ……... 43

Figure 10. Tandem simple repeat analysis of RPA ChIP-Seq samples ……... 46

Figure 11. Tandem simple repeat enrichment in RPA ChIP samples ……. 47

Figure 12. Non-contiguous simple repeat analysis of RPA ChIP-Seq samples ……. 49

Figure 13. Comparison of repeat length in RPLs to repeat length in the genome ……. 50

Figure 14. Non-denaturing 12% PAGE gel of repeats listed in Table 4 ……. 53

Figure 15. CD molar ellipticity of RPL repeat sequences ……. 55

Figure 16. CD wavelength scans normalized per CAGAGG repeat ……... 57

Figure 17. Representative CD melting curves for $(CAGAGG)_n$ ……..…... 58

Figure 18. Graph of melting temperatures obtained in UV-vis melting studies at different monomer lengths of the CAGAGG repeat ……..…... 58

Figure 19. Non-denaturing 12% PAGE gel of CA5, CA10, and CA15 …….……... 59

Figure 20. Overlay of normalized UV-vis melting data for $(CAGAGG)_{10}$ at varying strand concentrations ……..….. 60

Figure 21. CD scans at 4°C for $(CAGAGG)_{10}$ at varying strand concentrations post UV-vis melting ……..……..… 61

Figure 22. Schematic of *in vitro* primer extension assay …….……..….. 62

Figure 23. Representative images of Pol δHE reaction products …..…….. 63

Figure 24. Pol δHE termination probability ……..…..… 64

Figure 25. 2D gel of replication intermediates arising from replication through $(CAGAGG)_{105}$ in ori-proximal vector …….……... 65

Figure 26. Schematic of replication through ori-proximal vectors …… 67

Figure 27. Quantitation of the RFB index after ori-proximal vector replication in U2OS cells …… 68

Figure 28. 2D gel of replication intermediates arising from replication through $(CAGAGG)_{105}$ in ori-distal vector ……... 69

Figure 29. Schematic of BrITL method ……..… 75

Figure 30. Experimental schematic of induced site-specific break at I-PpoI site …….. 77

Figure 31. BrITL qRT-PCR of genomic regions surrounding the I-PpoI site …….. 78

Figure 32. BrITL qRT-PCR of RPA-enriched sites centered around $(CAGAGG/CCTCTG)_n$, $(CAGG/CCTG)_n$, $(CACAG/CTGTG)_n$, and $(CAAAA/TTTTG)_n$ repeats ……... 80

Figure 33. BrITL qRT-PCR of RPA-enriched site centered around $(CACAG/CTGTG)_n$ repeats ……... 82

Figure 34. Overlap of BrITL and RPA peaks in ATRi+aph^{18hrs} condition …….. 84

Figure 35. BrITL peaks at inverted SINEs …… 88

Figure 36. Model for fork collapse at $(CAGAGG)_n$ repeats ……..... 101

Figure 37. Model for fork collapse at inverted retroelements ……... 102

CHAPTER 1: INTRODUCTION

Permanent damage to the cellular genome is constantly limited by evolutionarily conserved signaling pathways to protect against propagation of critical errors or cell death. Exposure of cells to UV light, irradiation, increased reactive oxygen species, or toxic chemicals can lead to altered DNA bases and disrupted DNA metabolic processes. Yet cells have fine-tuned a series of responses and checkpoints that sense and correct the ensuing DNA damage in an effective manner. The stage at which the genome is most vulnerable to errors is during its replication.

Dedicated to efficiently and correctly duplicating the 3 billion base pairs of DNA, the S-phase of the cell cycle is equipped with resources that mitigate damage to the genome. Should damage occur, it is this part of the cell-cycle, in addition to G2, that promotes repair by homologous recombination (HR), a process that preserves accurate DNA sequences with no loss, in contrast to the error-prone repair mechanism mediated by non-homologous end-joining (NHEJ). During S-phase, replication forks that encounter difficult-to-replicate regions of the genome, such as extensive repeat-containing sites, or that confront DNA lesions, such as damaged DNA bases or ssDNA gaps, induce polymerase stalling and arrest DNA synthesis. These stalled forks, if not properly stabilized, can become susceptible to complexes that cleave the region into double-strand breaks (DSBs). To segregate chromosomes properly in cell division and to ensure correct replication of the cell's genetic material in S-phase, the DNA replication checkpoint response is activated to protect against DSB formation. This allows forks to resume replication normally and finish DNA synthesis correctly.

1.1 Replication stress

Replication stress is characterized by an increase in the frequency of stalled replication forks during DNA synthesis. Evidence of endogenous sources of replication stress comes from fundamental studies in yeast genomes, which have identified replication barriers, termed ‘slow zones’, that lead to chromosomal rearrangements under conditions of replication stress (Cha and Klecker, 2002). Examples of these barriers to replication include protein-bound DNA, secondary structure-forming DNA, and RNA:DNA hybrids, or R-loops (Figure 1). The presence of these structures on a strand can compromise replication fork progression and lead to fork stalling that, in the absence of proper stabilization and repair, may result in replication fork collapse and double-strand break formation, which serves as catalysts to rearrangements and mutations (Aguilera and Garcia-Muse, 2013). Various experiments have delineated several potential classes of innately susceptible sites in the mammalian genome that may play a crucial role in genomic instability associated with cancer and various other diseases. These are known as difficult-to-replicate loci.

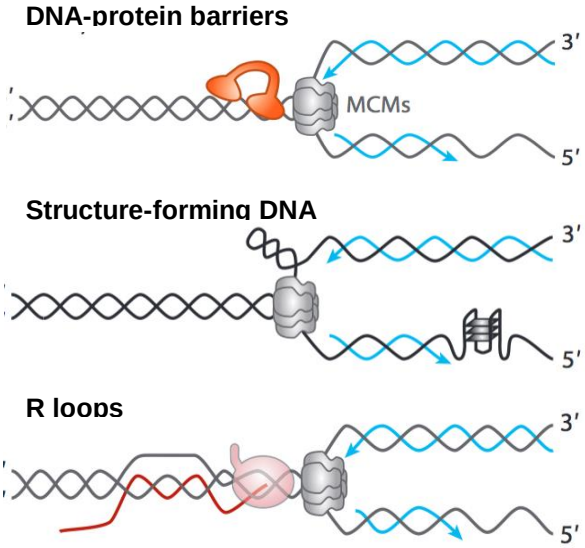

Figure 1. Hotspots for replication-induced rearrangements (Aguilera and Garcia-Muse, 2013).

Common Fragile Sites

The first class of these difficult-to-replicate loci consists of common fragile sites (CFS). These are defined as genomic regions that display gaps and breaks on metaphase chromosomes upon induction of mild replication stress, including treatment with low-dose aphidicolin, or inhibition of ATR (Casper et al., 2002). Instability at only 20 fragile sites comprises over 80% of all DNA lesions in lymphocytes following treatment with low-dose aphidicolin (Glover et al., 1984). In the absence of any concrete defining features, such as the presence of specific repeats (although these regions tend to be AT-rich) or altered forms of DNA structure, CFS nevertheless exhibit consistent and frequent events of sister chromatid exchanges, translocations, and deletions (Glover and Stein, 1987, 1988; Wang et al., 1997). Largely characterized in lymphocytes, top CFS include *FRA3B* at 3p14.2 and *FRA16D* at 16q23. Common traits of these loci are that they include spans of megabases, overlap with very large genes (>300 kb), and replicate in late S-phase (Helmrich et al., 2011; Smith et al., 2007).

Recently, a paucity of replication origins and limited replication initiation was demonstrated within a CFS to contribute to its fragility (Letessier et al., 2011). Furthermore, as distinctive cell types exhibit different replication origins and timing, the fragility of these sites can vary between cell types (Letessier et al., 2011). Since replication at these regions occurs late in S-phase, the ability to respond to fork slowing before completion of S-phase is compromised due to a dearth of nearby origins to fire and finish replication. Therefore, these sites are likely to enter mitosis while still under-replicated,

and to become processed as gaps and breaks on condensed metaphase chromosomes. Most interestingly, CFS are frequent locations of deletions and chromosome rearrangements in cancer cells, suggesting a key role for these loci in early genomic instability that contributes to cancer (Smith et al., 2007; Bignell et al., 2010).

DNA repeats

A second class of difficult-to-replicate loci is comprised of DNA repeats. Within this class are trinucleotide repeats (TNRs), long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs). Repetitive regions are pervasive in the genome, serving regulatory functions and supporting specific chromatin architecture. However, due to their repetitive nature, these sites become increasingly vulnerable to errors under conditions of reduced replicative efficiency. Tandem TNRs are susceptible to expansions and deletions during replication as a result of replication slippage or the presence of ssDNA that may allow secondary structure formation (Pearson et al., 2005). Studies in yeast and human cells have indicated that long tandem repeats can cause fork stalling as well as increased expansions and contractions in cells that are deficient in functional replicative components, further implicating the replication process in their instability (Aguilera and Garcia-Muse, 2013). Expansion of TNRs beyond a critical length (>35-40 units for CAG repeats; >200 units for CGG repeats; >100 units for GAA repeats) (Lahiri et al., 2004; Fu et al., 1991; Chutake et al., 2014) has been linked to genetic instability associated with several diseases, including Huntington's, Fragile X syndrome, and Friedreich's ataxia (Campuzano et al., 1996; Fu et al., 1991; Mandel and Heitz, 1992; Yudkin et al., 2014; Gerhardt et al., 2016).

Retroelements such as SINEs and LINEs are liable to generate genomic instability via a different mechanism. These retrotransposons are normally transcriptionally repressed, but become increasingly activated upon induction of cellular stress, leading to greater mobility and insertions at vulnerable sites in the genome. Copy number and mRNA levels of both LINE-1 and SINE B1 elements have been observed to increase upon onset of tumor formation in transgenic breast-cancer mice, which escalates further at later stages of tumor progression (Gualtieri et al., 2013). This indicates that repetitive sequences in the genome can change the genetic landscape by expanding, contracting, or inserting themselves into accessible regions because of the enhanced difficulty in processive and accurate replication through these sequences.

Non-B DNA

Lastly, the third class of difficult-to-replicate sites includes sequences that form non-B DNA structures. Beyond tandem repeats, this class encompasses motifs in DNA sequences that induce formation of specific structures, such as G-quadruplexes, hairpins or altered duplex forms of DNA. These secondary structures have the potential to form during replication when duplex strands become separated or energy is provided by negative supercoiling. Stable structure formation has been proven to impede fork progression and to serve as substrate for structure-specific nucleases that lead to DSB formation (Lobachev et al., 2002; Lopes et al., 2011). G-quadruplexes are generated by intra- or inter-molecular interactions between G-rich sequences on one or both DNA strands. Stabilization of these well-characterized quadruplex forms has been shown to slow the replication fork and lead to DSB formation and deletions at these regions (Paeschke, K et al., 2013; Bochman, M et al., 2012).

Similarly, hairpins and cruciforms, when produced, present a strong barrier to replicative polymerases that can lead to fork stalling. Specifically, long AT-rich palindromic regions on human chromosome 11q23 and 22q11 are frequent sites of breakpoint and translocation events in lymphoblasts and fibroblasts (Kurahashi, H et al., 2000). Furthermore, ectopic insertion of palindromic *Alu* sequences in yeast and mammalian cells has been shown to lead to fork stalling that is dependent upon hairpin formation from the sequence (Voineagu et al., 2008). Short inverted repeats, with their capability to form hairpin structures, stimulate breakage and deletion in cells and are enriched at translocation breakpoint junctions in cancers, demonstrating a role for these structures in cancer genomic instability (Lu et al., 2015). Besides stalling replication forks, non-B DNA structures can also decrease repair efficiency due to their altered forms and by preventing access to important repair proteins. Overall, sequences that have the opportunity to form non-B DNA structure in instances of ssDNA exposure, such as during replication, can hinder fork progression and lead to prolonged fork stalling.

In aggregate, these classes of genomic loci have been well-characterized to play a vital role in cancer and other disease states, suggesting that intrinsic DNA sequence and structure can be powerful sources of genomic instability due to their difficulties in replication. However, the contributions of each class of difficult-to-replicate loci under specific contexts of genomic instability remain insufficiently studied. Particularly, the relative vulnerabilities of different classes of sequences and the sensitizing or dampening role of local chromatin features have not been characterized comprehensively.

1.2 Replication checkpoint response

The induction of replication stress, either by exogenous damaging agents or by faulty progression through difficult-to-replicate loci, as described above, initiates a cascade of protective events to deter propagation of resultant errors in DNA synthesis. Impediments to polymerase progression across a DNA strand can lead to the uncoupling between the polymerase and the continuously unwinding minichromosome maintenance (MCM) helicase. This results in exposure of ssDNA between the polymerase and helicase that becomes coated by a single-stranded DNA binding protein, replication protein A (RPA). RPA-coated ssDNA recruits the ataxia telangiectasia and Rad3-related protein (ATR), a checkpoint kinase that is a key upstream regulator of the replication stress response pathway (Zou and Elledge, 2003; Flynn and Zou, 2011).

Along with its obligate partner, ATRIP, ATR becomes activated to stabilize the stalled forks by instigating a series of downstream events that promote cell cycle arrest and fork restart through phosphorylation of effector proteins, predominant of which is CHK1 (Capasso et al., 2002; Cimprich and Cortez, 2008; Cortez et al., 2001; Zhao and Piwnica-Worms, 2001). Once activated, CHK1 phosphorylates CDC25A, which causes its degradation and prevents its interaction with Cyclin A-Cdk1, Cyclin B-Cdk1, and Cyclin E-Cdk2; this inhibits origins from firing and arrests the cell cycle (Jin et al., 2003; Busino et al., 2003). CHK1 also phosphorylates CDC25C, which isolates it in the cytoplasm to prevent its activation of Cyclin B-Cdk1 (Sanchez et al., 1997). This halts cell cycle progression into M-phase until replication forks can restart (Liu et al., 2000). The combination of these events allows the cell to respond to the polymerase-stalling lesion and to resume replication normally within S-phase.

ATR also promotes protection of stalled forks through regulation of other substrates, such as SMARCAL1. SMARCAL1 is a DNA annealing helicase that stimulates fork-reversal of nascent DNA at stalled forks. Its phosphorylation by ATR limits its activity to minimize aberrant fork structures that are substrates for SLX4-mediated cleavage (Couch et al., 2013). ATR also promotes fork protection through regulation of repair factors such as WRN, a RecQ helicase that aids in fork restart and prevents DSBs, possibly by processing stalled fork intermediates and promoting recombination (Ammazzalorso et al., 2010).

1.3 Replication fork collapse and recovery

Fork collapse occurs when stalled replication forks are unable to resume replication, such as through loss of replisome components. Such events become more likely in the absence of fork-stabilizing functions downstream of ATR signaling. Inappropriate CDK activation in ATR-deleted cells appears to promote replisome disassembly; however, these findings may be context-dependent that is incompletely defined (Ragland et al., 2013; Cobb et al., 2003; De Piccoli et al., 2012; Dungrawala et al., 2015). In the absence of ATR, aberrantly activated CDK-dependent pathways promote the premature assembly of endonuclease complexes, such as SLX4-SLX1 or MUS81-EME1, that cleave replication forks into DSBs (Fekairi et al., 2009; Sarbajna et al., 2014; Pepe and West, 2014; Ragland et al., 2013; Szakal and Branzei, 2013; Couch et al., 2013). DSBs are lethal to a cell, and thus repair pathways become activated to repair them efficiently.

Formation of DSBs at forks induces ATR-mediated phosphorylation of the histone variant, H2AX, into γH2AX, which acts as a local marker of DNA DSBs and promotes the accumulation of various repair factors to the site of damage (Capasso et al., 2002; Brown

and Baltimore, 2003; Chanoux et al., 2009). While its complete function is still unknown, γH2AX is critical for stabilizing the interactions of factors necessary for the repair of the double-strand break (Rogakou et al., 1998; Rogakou et al., 1999; Furuta et al., 2003; Chanoux et al., 2009). The spread of γH2AX, which occurs up to several hundred kilobases on either side of the break point, restructures the chromatin to concentrate repair proteins to the site of damage (Rogakou et al., 1998; Rogakou et al., 1999; Coster and Goldberg, 2010). This includes components of repair involved in homologous recombination that re-establish the replication fork.

Phosphorylated H2AX recruits MDC1, which brings in the E3 ubiquitin ligase, RNF8 (Coster and Goldberg, 2010). RNF8 catalyzes the addition of ubiquitin molecules onto H2AX, which accumulates RNF168 and leads to the addition of poly-ubiquitination chains onto H2AX. These ubiquitin chains subsequently recruit 53BP1, the BRCA1-Abraxas-RAP80 complex, and MRN (Mre11-Rad50-Nbs1) to the break site (Jasin and Rothstein, 2013). Initial 5'-3' resection of the broken DNA end occurs by Mre11 and CtIP, and then more extensively by Exo1 and Dna1 nucleases. This allows RPA to bind to the stretch of ssDNA on the resected strand, with subsequent BRCA2-mediated formation of Rad51 filaments to replace RPA on the resected strands. The Rad51-coated strand then invades the nearby complimentary sister chromatid to utilize the homologous template for synthesis.

In the context of a one-ended break resulting from cleavage at a collapsed replication fork, break-induced replication (BIR) occurs (Constantino et al., 2014). This repair process involves the continued synthesis of the invading strand using the sister chromatid as a template and the catalytic subunit of DNA polymerase δ to finish synthesis to the end of the chromosome (Constantino et al., 2014). In this way, a collapsed fork is

able to recover and restart replication through homologous recombination-mediated repair. Through an overall integrated response, ATR stabilizes stalled forks to prevent double-strand break formation while phosphorylation of H2AX stimulates fork restart through homologous recombination if collapse occurs.

1.4 Cancer genomic instability

Cancer has often been described as a disease of the genome, characterized predominantly by abnormal growth and altered metabolism. These features break the boundaries of regular cell maintenance and allow the cells to proliferate beyond the normal capacity of untransformed cells. In this manner, a critical balance is shifted and homeostasis is lost, leading to a catastrophic state of cancer initiation and progression. Increased chromosomal instability and improper repair lead to greater levels of mutations that enhance a cancer cell's ability to evolve to the point of uncontrolled growth and drug resistance. How a cell transforms to achieve these features has been under investigation for decades. What is currently understood is that activation of oncogenes leads to abnormal growth in which cells cycle rapidly into and through S phase. Under this accelerated state of proliferation, DNA repair pathways can become less effective against increasing damage accrued during DNA synthesis, leading to replication stress. This causes DNA replication to become more reliant on DNA damage checkpoint responses to maintain genome stability. However, mutations may arise that bypass normally activated checkpoints, allowing cells that would generally senesce to survive under conditions of genomic instability. Increased mutagenesis and complex rearrangements that are advantageous to a cell's survival then promote clonal outgrowth and expansion to the detriment of surrounding normal cells.

Characterizing a cell's genomic landscape by identifying precisely where DNA breaks occur in the genome under conditions that cause the cell to propagate in a deregulated fashion will increase our understanding of cancer etiology. Similarly, such characterization under conditions that cause a cell to inherit a proclivity for certain diseases will expand our understanding of the mechanisms of genetic instability that allow the cell to accrue such damage and display the disease phenotype. While cancer genomic studies have identified translocation breakpoints and sites of large deletions, these samples were extracted from a static state of the cancer at a defined stage of its progression. DSB landscapes from these samples could prove to be different from tumor samples isolated at earlier stages of cancer initiation and progression. Because of our incomplete knowledge of genomic features that make certain chromatin regions more vulnerable to replication stress than others, the study of this question under these different contexts becomes even more relevant. A proper DNA break-detection assay is thus necessary to address such questions.

1.5 Break-identification methods

While it is possible to probe for the genomic location of DNA repair proteins to determine sites of DNA damage, it is but a substitute to query for the actual site of breakage. Furthermore, spreading of damage sensors and repair factors, such as γH2AX, may extend to kilobases and even megabases from the actual site of damage (Rogakou et al., 1998; Rogakou et al., 1999). This decreases precise characterization of the break site and increases background. Just recently have methods to detect DNA breaks arisen to address limitations previously imposed by low resolution and signal-to-noise ratio.

LM-PCR

One of the first methods developed was ligation-mediated PCR, or LM-PCR (Mueller et al., 2001). In this assay, cleaved DNA is denatured and hybridized to a gene-specific primer that extends DNA synthesis to the cleaved site, creating a blunt end. To this end, a linker is ligated. The DNA fragment is denatured again and hybridized to a second gene-specific primer that extends DNA synthesis to the linker sequence. PCR then occurs on these fragments with a linker sequence-specific primer and the second gene-specific primer. Doing so amplifies the target DNA, which is then sequenced and mapped to the genome. However, a major limitation to this protocol is that it requires prior knowledge about the break-proximal genes to recognize the presence of DSBs within the region. Thus, it cannot identify de novo DNA breaks.

HTGTS-Seq

High-throughput, genome-wide translocation sequencing, or HTGTS-Seq, involves a fixed 'bait' DSB that is created by a nuclease at an ectopically inserted recognition site in the genome (Chiarle et al., 2011). The 'bait' DSB is joined in a translocation event by an endogenously generated DSB. Genomic DNA is sonicated, ends polished, and adaptors are ligated to allow for PCR amplification through adaptor-specific primers. A 5' biotinylated primer complementary to the bait sequence synthesizes across the translocation junction with subsequent selection of the site by streptavidin binding. The junction is verified by sequencing. However, the efficiency of the assay is biased by the proximity of DNA ends for translocation and is only able to detect DSBs that translocate. It is possible that breaks that do not rejoin at all, but that persist, are not detected. Translocation events are rare and this method may thus necessitate a large amount of input DNA to be able to detect breaks. Because of these constraints, the assay is not quantitative.

BLESS

Breaks labeling, enrichment on streptavidin, and next-generation sequencing, or BLESS, aims to map DSBs at nucleotide resolution (Crosetto et al., 2013). In this protocol, cells are harvested and fixed. After sequential lysing of the cell and nuclear membranes, cells undergo a series of washes before resuspension in a reaction containing T4 DNA polymerase and T4 polynucleotide kinase (PNK) to blunt DNA ends. On the second day, cells are incubated overnight in a reaction mixture that ligates DNA ends with a linker. DNA is then extracted from the nuclear pellets and purified before being digested with the HaeIII restriction enzyme. Cleaved genomic DNA is subsequently sheared to smaller size fragments by sonication, and biotinylated DNA is captured through streptavidin-coated beads. Once attached to the beads, DNA ends are blunted and ligated to a second linker. The DNA is then eluted by digestion with I-SceI enzyme and PCR amplified for sequencing. A major setback to this procedure is that it utilizes fixed cells, which generates considerable breaks on its own, thus resulting in high-level background and low sensitivity.

Break-Seq

This assay was developed in Wenyi Feng's laboratory (Hoffman et al., 2015). It utilizes cells embedded in an agarose plug that are saturated in a reaction mixture containing T4 DNA polymerase to catalyze the addition of biotin-14-dATP to DNA ends through an end-repair mechanism. Chromatin is then purified from the plug, sheared to less than 500 bp, and captured through streptavidin-coated beads. Labeled DNA fragments attached to the beads are ligated to adaptors and amplified by PCR. The purified PCR products subsequently undergo paired-end sequencing by Illumina HiSeq. While it aims to discover de novo breaks in an unbiased manner, this protocol has been known to yield high

background. The signal-to-noise ratio of aligned tracks is typically low and does not include input normalization. Sensitivity remains an issue as signal from lower-amplitude breaks in this assay would likely get masked by high background.

GUIDE-Seq

Genome-wide, unbiased identification of DSBs enabled by sequencing, or GUIDE-seq, utilizes DNA repair pathways, such as non-homologous end joining (NHEJ), to mediate the integration of double-stranded oligonucleotides into DSB gaps (Tsai et al., 2015). Genomic DNA is sheared into ~500 bp and the ends are ligated to an adapter. These regions are then amplified by PCR utilizing the known sequences of the double-stranded oligonucleotides along with an additional nested PCR to gain high specificity of the target regions. However, the accuracy of this method depends upon the repair efficiency of integrated oligonucleotides at the break site, which may vary depending upon chromatin structure and sequence. Additionally, only breaks with blunted ends are identified since the double-stranded oligonucleotides are blunt-ended. Thus, this method is not quantitative, lending itself to biases towards accessible sites for integration and DSB ends without 5' or 3' overhangs.

End-Seq

End-sequencing is a method recently developed by Andre Nussenzweig's laboratory to capture DNA ends (Canela et al., 2016). It is the most sensitive assay to detect break sites genome-wide to date. However, it has only yet been studied in the context of RAG-mediated cleavage in pre-B cells and in thymocytes actively undergoing V(D)J recombination in G1 cells. In the protocol, at least 15 million cells are harvested and embedded in agarose plugs, washed and treated with Proteinase K and RNase A before

subsequently being stored at 4°C for up to 2 weeks. Although the purpose of labeling double-strand breaks in embedded plugs is to minimize creation of artefactual breaks, a high background may result. This was observed in our experience, presumably by passive depurination and depyrimidation events that take place and that are prevented from repair during the time cells are stored in plugs. In End-Seq, further processing occurs within these plugs: blunting, A-tailing, and ligation of DNA-damage ends to biotinylated hairpin adaptor 1. DNA is then purified from the plugs and sheared to 150-200 bp by sonication. Biotinylated DNA fragments are isolated by incubation with streptavidin beads and end-repaired for ligation to a second hairpin adaptor in preparation for library generation and deep sequencing. While this method proposes to detect breaks with nucleotide-level resolution, reads generated from sequences starting at the same base pair can produce significant amounts of duplicated reads that may incur false-positive peaks without a proper means of filtering out false duplicates from real duplicates. In addition, this method has not been proven to work in S-phase cells.

1.6 Significance of defining genomic sites that rely on ATR for stability

In general, DNA breaks create genomic instability, as they often lead to rearrangements and mutations that promote disease or cancer. Identification of these break sites through an unbiased and precise method is currently lacking. As ATR is a critical mediator that preserves genomic stability during replication, and as deletion of ATR leads to enhanced DNA breaks under conditions of replication stress and oncogene activation (Schoppy et al., 2012), a study to properly define the landscape and mechanisms of DNA breakage in the absence of ATR's activity will be critical.

Evidence exists to suggest that loss of ATR is sufficient to lead to DSBs. Inactivation of ATR by ATR deletion, shRNA-mediated gene suppression, and ATR kinase inhibition leads to an increase in chromosome breaks even in the absence of exogenous treatments that stall DNA replication (Brown and Baltimore, 2000; Brown and Baltimore, 2003; Chanoux et al., 2009). These breaks can form either stochastically across the genome during DNA synthesis, or occur preferentially at sites of slowed DNA synthesis. Suggestive of the latter, breakage at CFS is increased in ATR-suppressed cells. However, recent evidence indicates that this increase best correlates with the unimpeded progression of ATR inhibited cells into G2-M phase and the inadvertent cleavage of forks at sites of incompletely replicated DNA (Ying et al., 2013; Letessier et al., 2011). Indeed, DNA replication fork progression does not appear to be substantially slowed at CFS, which complete DNA replication late in S-G2 due to a greater distance between the early firing origins that flank these instability sites (Letessier et al., 2011). More recently it has been shown that the full assembly of the SLX4-1 endonuclease complex, which is at least partly responsible for DSB generation upon ATR inhibition (Ragland et al., 2013; Couch et al., 2013), occurs at the G2-M phase transition. As of yet, no study to define the landscape and mechanisms of DNA breakage in the absence of ATR's activity has been performed.

Data from yeast models suggests that under ATR inhibition some regions of the genome are more prone to replication-induced DNA damage than others (replication slow zones) (Cha and Klecker, 2002). But while certain replication stress-vulnerable sites are known based on yeast genetics and mammalian cytogenetics, there is a lack of knowledge on regions of the mammalian genome that break preferentially when the replication stress responses are defective. There is therefore a great need to accurately determine regions

that are prone to break in the absence of a replication checkpoint response and to understand why these sites break.

1.7 Aims

Based upon observations of replication-induced rearrangements and genomic instability that stem from specific regions sensitive to replication stress, I hypothesized that replication stress is more detrimental to certain regions of the genome than others. These would be revealed as defined fork collapse sites in the mouse genome. Furthermore, these sites would be predicted to occur at difficult-to-replicate regions, specifically at repetitive and structure-forming DNA.

To test this hypothesis, we aimed to capture and characterize highly frequent fork-collapse sites under specific conditions of replication stress. In our model, mouse embryonic fibroblasts ($ATR^{+/-}$) were treated with low-dose aphidicolin to induce replicative stress, and an ATR inhibitor (ATRi) to prevent repair and restart of stalled forks. In the absence of ATR, forks that have encountered uncoupling between the MCM2-7 helicase and polymerase are not readily stabilized, leading to increased RPA-bound intermediates that are an outcome of either stalled forks that have not resolved or resected strands of DSBs that arise from the collapse of these stalled forks. Sites of persistent fork collapse in the mouse genome were subsequently isolated and mapped by two methods: RPA ChIP-Seq, and a novel DNA break-detection assay, BrITL.

Using RPA ChIP-Seq and BrITL, I proposed to identify the locations in the genome that are rendered most sensitive to ATR inhibition when replication stress is applied and to determine how DNA sequence at replication-sensitive loci contributes to replication-induced instability. By characterizing genomic regions that most frequently lead to DSBs

in the absence of a functional ATR repair pathway under replication stress, I plan to address how unchecked replication stress can promote distinctive sites of genomic instability. In doing so, I have been able to identify possible mechanisms by which decreased genome surveillance can transform a cell under specific stress conditions.

In my first aim, I will describe findings on identified and enriched repeats within RPA ChIP-Seq peaks specific to the combination treatment of ATRi and aphidicolin (ATRi+aph) utilizing a developed repeat analysis program called REQer (Repeat Enrichment Quantifier). From deeply sequenced data generated by RPA ChIP-Seq, REQer allowed us to dissect the role of repetitive regions in replication stress with greater depth. While DNA repeats are difficult to quantify through normal ChIP-Seq metrics due to their non-unique mappability, through REQer, we have developed a method to characterize the abundance of repetitive elements in the ChIP versus input samples. The most enriched repetitive sequences were further tested for their ability to cause polymerase pausing both *in vivo* and *in vitro* and were queried for possible formation of secondary structures, addressing mechanisms by which these sequences may lead to RPA accumulation and fork collapse under replication stress.

In my second aim, I will describe the development and use of a break-detection assay termed BrITL (Break Identification by TdT Labeling) to confirm that fork-collapse sites identified by RPA ChIP-Seq are consequently sites of frequent DNA breakage. By this study, two categories of sites were discovered in replication-stressed mouse embryonic fibroblasts: 1) the overlap of peaks from RPA ChIP-Seq and BrITL, and 2) exclusive BrITL peaks that are not sites of significant RPA accumulation. Due to their identification in both RPA ChIP-Seq and BrITL, the sites in category 1 were defined as fork collapse regions that degenerate frequently into double-strand breaks. The subset of

breaks described in category 2 suggests sites that, by a different mechanism, do not recruit significant amounts of RPA, but still lead to frequent breakage. Surprisingly, this subset of breaks was found to associate with inverted sequences composed entirely of SINE, LINE or LTR retroelements. Structural data analysis indicated that these sites form highly stable hairpin structures that preclude RPA accumulation.

Herein, I report identification of 173 Replication Perturbed Locations (RPLs) across the mouse genome caused by ATR inhibition employing RPA ChIP-Seq. These sites were enriched for specific simple repeats, one of which is implicated in myotonic dystrophy type II (CAGG/CCTG), but most that have not been previously characterized as difficult to replicate. Furthermore, triplet repeats commonly associated with disorders in their expanded states, described previously (GAA/TTC, CAG/CTG, CGG/CCG), were not found to be significantly enriched in RPL sites. Structural assays on the most common RPA-associated simple repeat (CAGAGG/CCTCTG) suggested a mechanism of fork collapse that occurs through the formation of a stable secondary structure. Both *in vivo* and *in vitro* replication stalling assays confirmed that this simple repeat is sufficient to impede fork progression. Furthermore, a recently developed break-detection assay termed BrITL recognized consistent breakage occurring at RPL sites associated with CAGAGG/CCTCTG repeats from ATR inhibition. Moreover, genome-wide BrITL results identified a greater subset of breaks that were not associated with RPA accumulation. Instead, a majority of these DNA breaks were centered around SINE, LINE or LTR elements that were inverted, leading to predicted hairpin structures. This finding was novel and suggested a new class of sequences highly susceptible to DSBs upon ATR inhibition.

Generally, reliable methods to detect the location of frequently occurring double-strand breaks in the genome and to examine why they may arise in certain regions over

others would elucidate how DNA damage under replication stress can promote the transition of a normal cell into a diseased or cancerous state through a targeted and specific manner. These studies can provide a more comprehensive understanding of how the process of mutagenesis transpires as a cancer cell evolves, or how certain polymorphic sequences primes a cell for genetic instability, leading to more precise targeting of the abnormal cell's genome versus that of a normal cell.

References

Aguilera A, Garcia-Muse T. (2013). Causes of genome instability. Annu Rev Genet. *47*, 1-32.

Ammazzalorso F, Pirzio LM, Bignami M, Franchitto A, Pichierri P. (2010). ATR and ATM differently regulate WRN to prevent DSBs at stalled replication forks and promote replication fork recovery. EMBO J. *29*, 3156-3169.

Bochman ML, Paeschke K, Zakian VA. (2012). DNA secondary structures: stability and function of G-quadruplex structures. Nat rev Genet. *13*, 770-780.

Brown EJ, Baltimore D. (2000). ATR disruption leads to chromosomal fragmentation and early embryonic lethality. Genes Dev. *14*, 397-402.

Brown EJ, Baltimore D. (2003). Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. Genes Dev. *17*, 615-28.

Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C, Widaa S, Hinton J, Fahey C, Fu B, Swamy S, Dalgliesh GL, The BT, Deloukas P, Yang F, Campbell PJ, Futreal PA, Stratton MR. (2010). Signatures of mutation and selection in the cancer genome. Nature. *463*, 893-898.

Busino L, Donzelli M, Chiesa M, Guardavaccaro D, Ganoth D, Dorrello NV, Hershko A, Pagano M, Draetta GF. (2003). Degradation of Cdc25A by beta-TrCP during S phase and in response to DNA damage. Nature. *426*, 87-91.

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Cañizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, De Michele G, Filla A, De Frutos R, Palau F, Patel PI, Di Donato S, Mandel JL, Cocozza S, Koenig M, Pandolfo M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. *271*, 1423-1427.

Canela A, Sridharan S, Sciascia N, Tubbs A, Meltzer P, Sleckman BP, Nussenzweig A. (2016). DNA breaks and end resection measured genome-wide by end-sequencing. Mol Cell. *63*, 898-911.

Capasso H, Palermo C, Wan S, Rao H, John UP, O'Connell MJ, Walworth NC. (2002). Phosphorylation activates Chk1 and is required for checkpoint-mediated cell cycle arrest. J Cell Sci. *115*, 4555-64.

Casper AM, Nghiem P, Arlt MF, Glover TW. (2002). ATR regulates fragile site stability. Cell. *111*, 779-789.

Cha RS, Klecker N. (2002). ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. Science. *297*, 602-606.

Chanoux RA, Yin B, Urtishak KA, Asare A, Bassing CH, Brown EJ. (2009). ATR and H2AX

cooperate in maintaining genome stability under replication stress. J. Biol. Chem. *284*, 5994-6003.

Chiarle R, Zhang Y, Frock RL, Lewis SM, Molinie B, Ho YJ, Myers DR, Choi VW, Compagno M, Malkin DJ, Neuberg D, Monti S, Giallourakis CC, Gostissa M, Alt FW. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell. *147*, 107-119.

Chutake YK, Lam C, Costello WN, Anderson M, Bidichandani SI. (2014). Epigenetic promoter silencing in Friedreich ataxia is dependent on repeat length. Ann Neurol. *76*, 522-528.

Cimprich KA, Cortez D. (2008). ATR: an essential regulator of genome integrity. Nat. Rev. Mol. Cell Biol. *9*, 616–627.

Cobb JA, Bjergbaek L, Shimada K, Frei C, Gasser SM. (2003). DNA polymerase stabilization at stalled replication forks requires Mec1 and the RecQ helicase Sgs1. EMBO J. *22*, 4325-4336.

Constantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Illiakis G, Kallioniemi OP, Halazonetis TD. (2014). Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science. *343*, 88-91.

Cortez D, Guntuku S, Qin J, Elledge SJ. (2001). ATR and ATRIP: partners in checkpoint signaling. Science. *294*, 1713–1716.

Coster G, Goldberg M. (2010). The cellular response to DNA damage: a focus on MDC1 and its interacting proteins. *1*, 166-178.

Couch FB, Bansbach CE, Driscoll R, Luzwick JW, Glick GG, Betous R, Carroll CM, Jung SY, Qin J, Cimprich KA, Cortez D. (2013). ATR phosphorylates SMARCL1 to prevent replication fork collapse. Genes Dev. *27*, 1610-1623.

Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, Karaca E, Chiarle R, Skrzypczak M, Ginalski K, Pasero P, Rowicka M, Dikic I. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods. *10*, 361-365.

De Piccoli G, Katou Y, Itoh T, Nakato R, Shirahige K, Labib K. (2012). Replisome stability at defective DNA replication forks is independent of S phase checkpoint kinases. Mol. Cell. *45*, 696-704.

Dungrawala H, Rose KL, Bhat KP, Mohni KN, Glick GG, Couch FB, Cortez D. (2015). The replication checkpoint prevents two types of fork collapse without regulation replisome stability. Mol. Cell. *59*, 998-1010.

Fekairi S, Scaglione S, Chahwan C, Taylor ER, Tissier A, Coulon S, Dong MQ, Ruse C, Yates JR 3rd, Russell P, Fuchs RP, McGowan CH, Gaillard PH. (2009). Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination

endonucleases. Cell. *138*, 78-89.

Flynn RL and Zou L. (2011). ATR: a master conductor of cellular responses to DNA replication stress. Trends Biochem Sci. *36*, 133-140.

Furuta T, Takemura H, Liao ZY, Aune GJ, Redon C, Sedelnikova OA, Pilch DR, Rogakou EP, Celeste A, Chen HT, Nussenzweig A, Aladjem MI, Bonner WM, Pommier Y. (2003). Phosphorylation of histone H2AX and activation of Mre11, Rad50, and Nbs1 in response to replication-dependent DNA double-strand breaks induced by mammalian DNA topoisomerase I cleavage complexes. J Biol Chem. *278*, 20303-12.

Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkert AJ, Holden JJ, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell. *67*, 1047-1058.

Gerhardt J, Bhalla AD, Butler JS, Puckett JW, Dervan PB, Rosenwaks Z, Napierala M. (2016). Stalled DNA replication forks at the endogenous GAA repeats drive repeat expansion in Friedreich's ataxia cells. Cell Rep. *16*, 1218-1227.

Glover TW, Berger C, Coyle J, Echo B. (1984). DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. Hum Genet. *67*, 136-142.

Glover TW, Stein CK. (1987). Induction of sister chromatid exchanges at common fragile sites. Am J Hum Genet. *41*, 882-890.

Glover TW, Stein CK. (1988). Chromosome breakage and recombination at fragile sites. Am J Hum Genet. *43*, 265-273.

Göhler T, Sabbioneda S, Green CM, Lehmann AR. (2011). ATR-mediated phosphorylation of DNA polymerase n is needed for efficient recovery from UV damage. J. Cell Biol. *192*, 219-227.

Gualtieri A, Andreola F, Sciamanna I, Sinibaldi-Vallebona P, Serafino A, Spadafora C. (2013). Increased expression and copy number amplification of LINE-1 and SINE B1 retrotransposable elements in murine mammary carcinoma progression. Oncotarget. *4*, 1882-1893.

Helmrich A, Ballarino M, Tora L. (2011). Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol Cell. *44,* 966-977.

Hoffman EA, McCulley A, Haarer B, Arnak R, Feng W. (2015). Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. Genome Res. *25*, 402-412.

Jasin M, Rothstein R. (2013). Repair of strand breaks by homologous recombination. Cold Spring Harb Perspect Biol. *5*, a012740.

Jin J, Shirogane T, Xu L, Nalepa G, Qin J, Elledge SJ, Harper JW. (2003). SFCbeta-TRCP links Chk1 signaling to the degradation of the Cdc25A protein phosphatase. Genes Dev. *17*, 3062-3074.

Kurahashi H, Shaikh TH, Hu P, Roe BA, Emanuel BS, Budarf ML. (2000). Regions of genomic instability on 22q11 and 11q23 as the etiology for the recurrent constitutional t(11;22). Hum Mol Genet. *9*, 1665-1670.

Kurahashi H, Shaikh TH, Emanuel BS. (2000). Alu-mediated PCR artifacts and the constitutional t(11;22) breakpoint. Hum Mol Genet. *9*, 2727-2732.

Lahiri M, Gustafson TL, Majors ER, Freudenreich CH. (2004). Expanded CAG repeats activate the DNA damage checkpoint pathway. Mol Cell. *15*, 287-293.

Letessier A, Millot GA, Koundrioukoff S, Lachages AM, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M. (2011). Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature. *470*, 120-123.

Liu Q, Guntuku S, Cui XS, Matsuoka S, Cortez D, Tamai K, Luo G, Carattini-Rivera S, DeMayo F, Bradley A, Donehower LA, Elledge SJ. (2000). Chk1 is an essential kinase that is regulated by ATR and required for the G2/M DNA damage checkpoint. Genes Dev. *14*, 1448–1459.

Lobachev KS, Gordenin DA, Resnick MA. (2002). The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell. *108*, 183-193.

Lopes J, le Piazza AE, Bermejo R, Kriegsman B, Colosio A, Teulade-Fichou MP, Foiani M, Nicolas A. (2011). G-quadruplex-induced instability during leading-strand replication. EMBO J. *30*, 4033-4046.

Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM. (2015). Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. Cell Rep. *S2211-1247*, 00197-7.

Mandel JL, Heitz D. (1992). Molecular genetics of the fragile-X syndrome: a novel type of unstable mutation. Curr Opin Genet Dev. *2*, 422-430.

Mueller PR, Wold B, Garrity PA. (2001). Ligation-mediated PCR for genomic sequencing and footprinting. Curr Protoc Mol Biol. *Chapter 15*, Unit 15.3.

Paeschke K, Bochman ML, Garcia PD, Cejka P, Friedman KL, Kowalczykowski SC, Zakian VA. (2013). Pif1 family helicases suppress genome instability at G-quadruplex motifs. Nature. *497*, 458-462.

Pearson CE, Nichol Edamura K, Cleary JD. (2005). Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet. *6*, 729-742.

Pepe A, West S. (2014). MUS81-EME2 promotes replication fork restart. Cell Reports. *7*, 1048-1055.

Ragland RL, Patel S, Rivard RS, Smith K, Peters AA, Bielinsky AK, Brown EJ. (2013). RNF4 and PLK1 are required for replication fork collapse in ATR-deficient cells. Genes Dev. *27*, 2259-2273.

Rogakou EP, Boon C, Redon C, Bonner WM. (1999). Megabase chromatin domains involved in DNA double-strand breaks in vivo. J Cell Biol. *146*, 905-916.

Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. (1998). DNA double-strand breaks induce histone H2AX phosphorylation on serine 139. J Biol Chem. *273*, 5858-5868.

Sanchez Y, Wong C, Thoma RS, Richman R, Wu Z, Piwnica-Worms H, Elledge SJ. (1997). Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25. Science. *277*, 1497-1501.

Sarbajna S, Davies D, West S. (2014). Roles of SLX1-SLX4, MUS81-EME1, and GEN1 in avoiding genome instability and mitotic catastrophe. Genes Dev. *28*, 1124-1136.

Schoppy DW, Ragland RL, Gilad O, Shastri N, Peters AA, Murga M, Fernandez-Capetillo, Diehl JA, Brown EJ. (2012). Oncogenic stress sensitizes murine cancers to hypomorphic suppression of ATR. J Clin Invest. *122*, 241-252.

Smith DI, McAvoy S, Zhu Y, Perez DS. (2007). Large common fragile site genes and cancer. Semin Cancer Biol. *17*, 31-41.

Szakal B, Branzei D. (2013). Premature Cdk1/Cdc5/Mus81 pathway activation induces aberrant replication and deleterious crossover. EMBO J. *32*, 1155-1167.

Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, Aryee MJ, Joung JK. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol. *33*, 187-197.

Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. (2008). Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. Proc Natl Acad Sci USA. *105*, 9936-9941.

Wang L, Paradee W, Mullins C, Shridhar R, Rosati R, Wilke CM, Glover TW, Smith DI. (1997). Aphidicolin-induced FRA3B breakpoints cluster in two distinct regions. Genomics. *41*, 485-488.

Ying S, Minocherhomji S, Chan KL, Palmai-Pallag T, Chu WK, Wass T, Mankouri HW, Liu Y, Hickson ID. (2013). MUS81 promotes common fragile site expression. Nat. Cell Biol. *15*, 1001-1007.

Yudkin D, Hayward BE, Aladjern MI, Kumari D, Usdin K. (2014). Chromosome fragility and the abnormal replication of the FMR1 locus in fragile X syndrome. Hum Mol Genet. *23*, 2940-2952.

Zhao H, Piwnica-Worms H. (2001). ATR-mediated checkpoint pathways regulate phosphorylation and activation of human Chk1. Mol. Cell. Biol. *21*, 4129–4139.

Zou L, Elledge SJ. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. Science. *300*, 1542–1548.

Contributions of work presented in Chapter 2:

- RPA ChIP-Seq was performed on MEFs by Dr. Yu-Chen Tsai (former Post-Doc, Brown Lab).
- REQer was developed by myself, Dr. Yu-Chen Tsai, and Dillon Maloney.
- Biophysical structural assays on the synthetic oligonucleotides and respective analyses were performed by Dr. Liliya Yatsunyk and students, Jessica Chen, Barrett Powell, and Deondre Jordan (Swarthmore, PA).
- *In vitro* primer extension assays, 2D gel electrophoresis, and respective analyses were performed by Drs. Kristin Eckert and Suzanne Hile (Penn State, PA) from subcloned vectors made by Dr. Yu-Chen Tsai.

CHAPTER 2: RPA CHIP-SEQ ON REPLICATION-STRESSED MEFs

2.1 Genome-wide identification of RPA-enriched sites caused by ATR inhibition

Unbiased identification of difficult-to-replicate sequences from ATR inhibition remains incomplete due in large part to the technological limitations of currently available assays. Early studies in yeast relied on phosphorylation of H2AX (γH2AX), which occurs up to and beyond 100 kb along chromatin following DSB generation from polymerase stalling in Mec1-Rad53 (ATR-CHK1 orthologue) suppressed cells (Szilard et al., 2010). These studies mainly identified regions of large tandem repeats, such as tRNA and rRNA gene arrays as sites of stalled replication. Other sites either were not detectable or were not found to be fragile in yeast under replication checkpoint abrogation. Other studies using higher-resolution DSB-detection methods, such as Break-Seq, identified different sites that were more closely associated with gene-rich regions (Hoffman et al., 2015). Finally, a study on mammalian cells characterized sites of fragility upon complete inhibition of fork progression at early DNA replication origins; in this case, however, ATR inhibition was only used to bypass the G2/M checkpoint to confirm breakage at these identified sites by analysis of mitotic spreads (Barlow et al., 2013). Others have used γH2AX to identify CFS or vulnerable regions of replication progression, but only from aphidicolin treatment without inhibition of ATR (Harrigan et al., 2011). Currently, no unbiased approach to detect sites of problematic replication with ATR inhibition has been reported.

In our study, replication stress is induced in passage-immortalized mouse embryonic fibroblasts ($ATR^{+/-}$) via treatment with ATR inhibitor (ATRi, 1 μM) accompanied by a partial inhibitory concentration of aphidicolin (aph, 0.2 μM) to enhance replication perturbation. At regions of enhanced fork slowing, uncoupling between the MCM2-7 helicase and polymerase occurs more readily, exposing stretches of ssDNA that become

bound by RPA. Accumulation of RPA-bound sites recruits the replication checkpoint kinase ATR to stabilize the stalled forks and to promote fork restart (Zou and Elledge, 2003; Cimprich and Cortez, 2008). However, replication intermediates arising from stalled forks that are normally transient will become persistent when the normal replication stress-recovery process mediated by ATR is inhibited, rendering stalled forks destabilized. Notably, ATR suppression has been shown to cause chromatid breaks even in the absence of exogenous stalling agents (Brown and Baltimore, 2000, Brown and Baltimore, 2003), indicating that some regions of the genome may be inherently difficult to replicate.

In this manner, ATR inhibition could be used as a tool to promote replication fork collapse specifically at these genomic sequences, which putatively stall polymerases and cause the accumulation of single-stranded DNA. Thus, in conditions where ATR is inhibited, increased instances of collapsed forks (sites that are no longer able to resume replication) will be available for capture (Figure 2). Accordingly, for our experiments, cells were treated with both aphidicolin and a specific ATR inhibitor, ATR-45 (Charrier et al., 2011). Under these conditions, RPA binds to sites of exposed ssDNA at destabilized forks resulting from both the uncoupled helicase and polymerase and the resected ends of forks that degenerate into DSBs (Figure 2). To map these proposed ATR inhibitor-sensitive sites, we applied RPA-chromatin immunoprecipitation (ChIP) followed by next-generation sequencing (NGS).

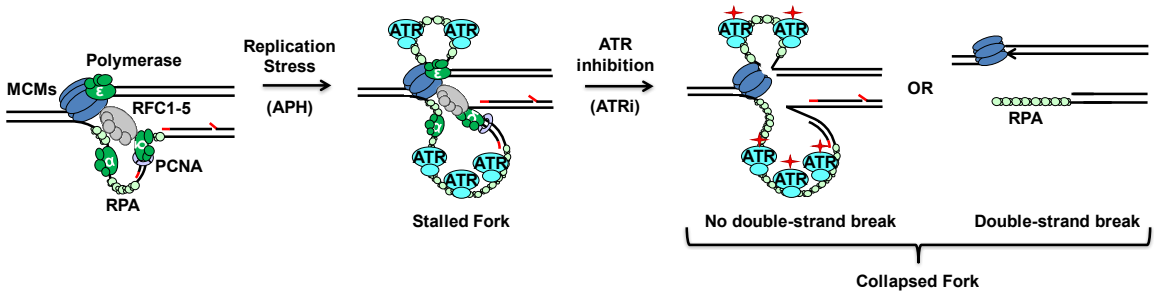

Figure 2. Schematic of replication stress-induced fork stalling. This scenario arises from low dose aphidicolin (APH) treatment and ATR inhibition through ATR inhibitor (ATRi), indicated by the red diamond.

Because tandem repeats have been implicated previously in fork stalling (Lahiri et al., 2004; Campuzano et al., 1996; Fu et al., 1991; Mandel and Heitz, 1992), we reasoned that the sonication fragment sizes typically used for NGS (200-300 bp) would be too short to contain non-repeat containing sequences that could be mapped to the reference genome. Therefore, RPA-coated chromatin was initially sonicated to 500-2,000 bps for RPA-ChIP retrieval, and retrieved DNA was then sonicated further to fragment sizes better suited for NGS (200-300 bp, Figure 3). This approach increased the likelihood that fork collapse due to repetitive DNA sequences could be mapped to the reference genome through unique sequences adjacent to the repeats. RPA ChIP-Seq was also performed on DMSO-only treated (UT) controls. Pre-ChIP input DNAs from each condition were isolated and sequenced as normalization controls.

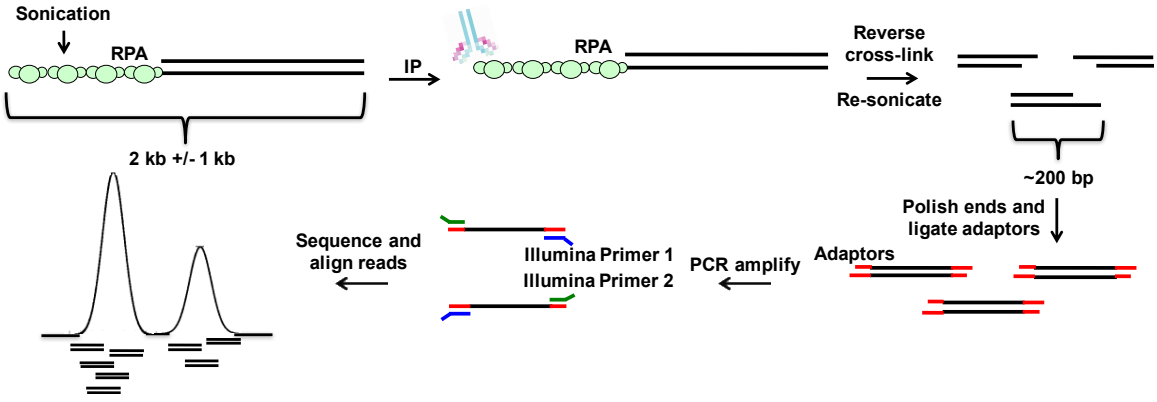

Figure 3. Schematic of RPA ChIP-Seq method. Cross-linked chromatin is sonicated into large fragments (<4 kb) and immunoprecipitated with an RPA-32 antibody. Samples are then reverse cross-linked and sonicated into smaller fragments (200-300 bp). DNA ends are polished, adaptor-ligated and PCR-amplified with Illumina primers into a library for

next-generation sequencing through the Illumina HiSeq platform. Reads generated from a 100 bp single-end sequencing run are aligned to the reference mouse genome.

Following alignment to the reference genome and normalization by input DNA of RPA ChIP-Seq reads from mouse fibroblast cells treated with either DMSO or 1 μM ATR-45 (ATRi) and 0.2 μM aphidicolin for 18 hours, loci characterized by statistically significant read enrichments (>4-fold over input, p-value $<10^{-3}$) in both of 2 biological replicates were identified. For enrichment analysis, the biological replicates and inputs of each experimental condition underwent an irreproducibility rate (IDR) analysis (Landt et al., 2012) with the MACS2 peak-calling program (Zhang et al., 2008) to give the final peak list per condition. IDR thresholds of >0.05 were used for self-consistency and comparison of biological replicates, and >0.005 for pooled-consistency analysis. Peaks that passed IDR thresholds were further filtered to select those with p-value $<10^{-3}$ and that were above 4-fold enriched over input. Regions within 2 kb of one another were merged. The final peak list per condition was generated as a set intersection with and subtraction from the DMSO-control peak list.

In total, 173 sites of significant and specific RPA enrichment were identified in the ATRi+aph^{18hrs} condition that were not observed in the DMSO-treated controls (Figure 4). Defined accumulations of reads within the coverage track of the ATRi+aph^{18hrs} RPA ChIP in Figure 4 represent peaks that are not observed in the coverage track of the DMSO-treated (UT) RPA ChIP. Peaks in the ATRi+aph^{18hrs} tracks were further normalized by both its own input and the untreated control, their sustained read enrichment indicating that these RPA-binding sites are specific to the treatment conditions and thereby representative of replication stress (Figure 4).

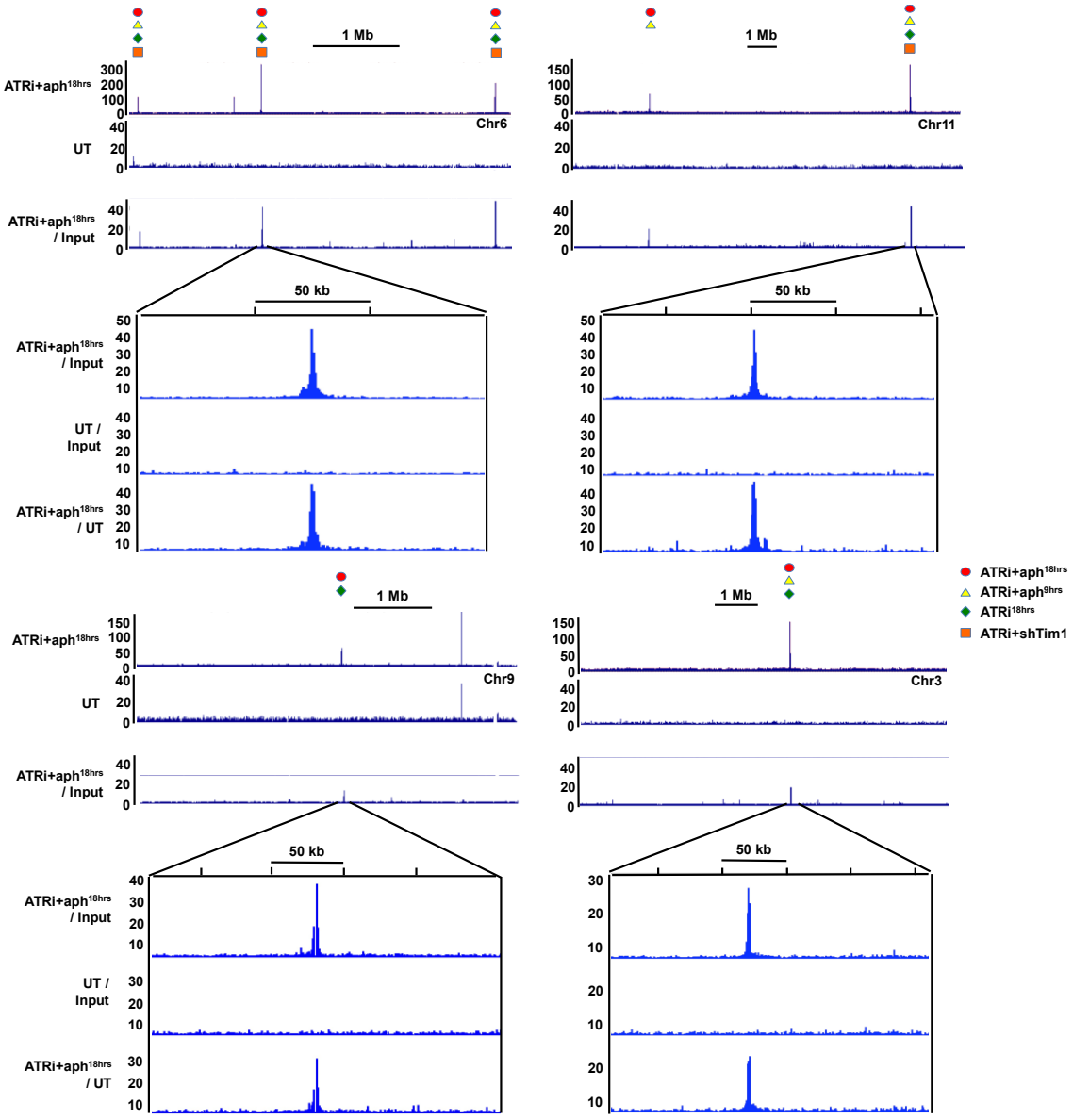

Figure 4. RPA ChIP-Seq coverage and ratio tracks on different chromosomes. Top row displays the coverage track of ATRi+aph^{18hrs} (RPA ChIP); the second row displays the coverage track of UT (RPA ChIP); the third row displays the ratio track of ATRi+aph^{18hrs} over its input (RPA ChIP/Input), using 500 bp bins. Symbols above select peaks indicate the identification of the peak under different experimental conditions. For zoomed-in tracks, the top row depicts the ratio track of combined replicates of ATRi+aph^{18hrs} over combined replicates of its input (RPA ChIP/Input); the second row depicts the ratio track of combined replicates of UT over combined replicates of its input (RPA ChIP/Input); the third row depicts the ratio track of combined replicates of ATRi+aph^{18hrs} over combined replicates of UT (RPA ChIP/RPA ChIP). Ratio tracks were generated using 750 bp bins.

To further refine and classify the RPA enrichment sites rendered vulnerable to replication fork collapse by ATR inhibition, three additional ATRi-treatment conditions were examined by RPA ChIP-Seq. These conditions included: 1) 9 hr treatment (ATRi+aph^{9hrs}), 2) ATRi in combination with suppression of the replisome factor TIMELESS (ATRi+shTIM1), and 3) ATRi treatment alone (ATRi^{18hrs}). TIMELESS is a protein in complex with TIPIN that has a role in fork protection, the down-regulation of which leads to replisome dysfunction and increased levels of ssDNA (Smith et al., 2009). As displayed in Figure 4, considerable peak overlaps were observed between different conditions of ATR-inhibited cells. Interestingly, few treatment-dependent differences were observed in the sites identified under these various conditions, and each of these sites was a subset of those identified by the ATRi+aph^{18hrs} condition (Figure 5). Of note, the single ATRi+shTIM1-specific peak presented some level of read accumulation in the coverage track of ATRi+aph^{18hrs}, however it was below the 4-fold cut-off. Similarly, two ATRi^{18hrs}-specific peaks presented some level of read accumulation in the ATRi+aph^{18hrs} condition, but did not pass peak thresholds for ATRi+aph^{18hrs}. The aggregate data sets from these independent experimental conditions were then used to create tiered categories of replication fork collapse sites (Figure 6).

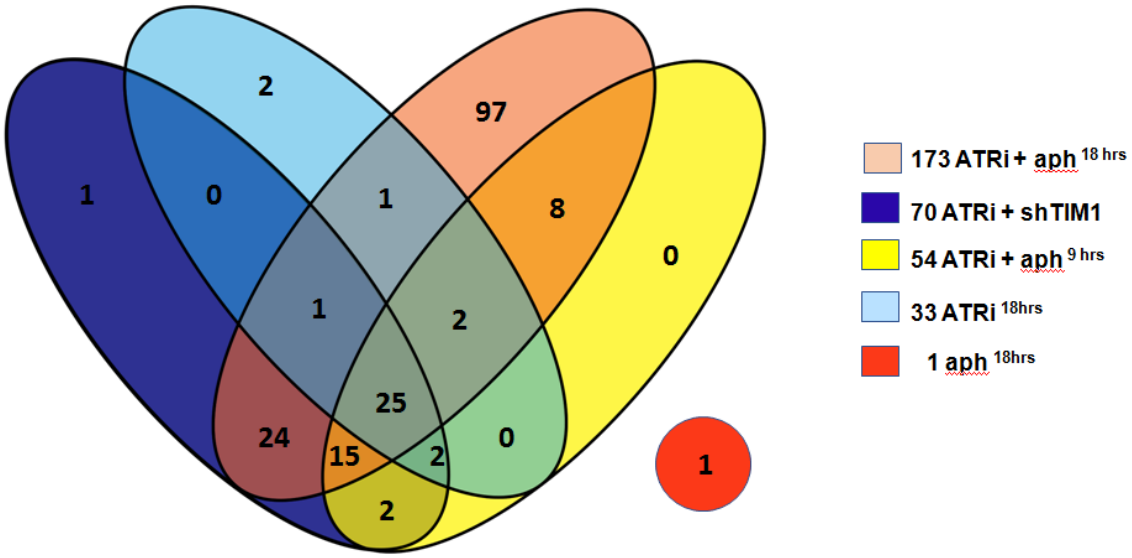

Figure 5. Venn diagram overlap of peaks identified from each ATRi condition.

Of the 173 sites identified in the ATRi+APH^{18hrs} condition (ATRi^{173}), a total of 55 peaks were observed in 2 or 3 additional conditions (ATRi^{55}), and 25 sites (ATRi^{25}) were identified in all conditions (Figure 6). We defined these ATRi-induced RPA ChIP-Seq sites as Replication Perturbed Locations, or RPLs, for brevity. Notably, the combined length of all 173 sites encompassed a small fraction of the total murine genome (<10^{-4}), indicating a high degree of specificity. This specificity, as well as our categorization of sites into tiers, permitted correlation of local genomic characteristics (repeat sequences, gene expression, and chromatin state) with the degree of ATRi-site vulnerability.

chrom	chromStart	chromEnd	ATRi+aph^{18hrs}	ATRi+aph^{9hrs}	ATRi^{18hrs}	ATRi+shTIM1
chr4	60081743	60081876	1	1	1	1
chr5	38023066	38023249	1	1	1	1
chr6	86329935	86334080	1	1	1	1
chr6	87771924	87779491	1	1	1	1
chr6	90513949	90522651	1	1	1	1
chr6	122612508	122622318	1	1	1	1
chr7	10285004	10285130	1	1	1	1
chr7	35155299	35164001	1	1	1	1
chr7	73320841	73326802	1	1	1	1
chr7	113703689	113703838	1	1	1	1
chr7	140221255	140221439	1	1	1	1
chr10	3110110	3117163	1	1	1	1
chr10	52221370	52221498	1	1	1	1
chr11	44500283	44505651	1	1	1	1
chr11	66421055	66421244	1	1	1	1
chr11	68502354	68502470	1	1	1	1
chr12	84280554	84285121	1	1	1	1
chr13	74528418	74531130	1	1	1	1
chr13	118058644	118058798	1	1	1	1
chr14	27528258	27528395	1	1	1	1
chr14	63277106	63283413	1	1	1	1
chr17	25451787	25451937	1	1	1	1
chrX	37130496	37130630	1	1	1	1
chrX	103749802	103749937	1	1	1	1
chrX	135048202	135054732	1	1	1	1

chrom	chromStart	chromEnd	ATRi+aph^{18hrs}	ATRi+aph^{9hrs}	ATRi^{18hrs}	ATRi+shTIM1
chr1	148723532	148723658	1	1	0	1
chr2	52694337	52700157	1	1	0	1
chr2	75747325	75749149	1	1	0	1
chr2	94944106	94944326	1	1	0	1
chr3	31069063	31072746	1	1	1	0
chr3	150557769	150557883	1	1	0	1
chr6	87457248	87458124	1	1	0	1
chr6	129096072	129096298	1	1	0	1
chr8	70664433	70664534	1	1	0	1
chr8	83857833	83861014	1	1	1	0
chr9	99705769	99705925	1	1	0	1
chr10	23988590	23988720	1	1	0	1
chr10	113920235	113920424	1	1	0	1
chr11	5740329	5743870	1	0	1	1
chr11	89081690	89082037	1	1	0	1
chr12	24768572	24768786	1	1	0	1
chr15	38308499	38308728	1	1	0	1
chr17	13305944	13308113	1	1	0	1
chr5	116363719	116363851	0	1	1	1
chr11	84152006	84152128	0	1	1	1
chr1	5071589	5072789	1	0	0	1
chr4	45193482	45194256	1	0	0	1
chr4	150644934	150645145	1	0	0	1
chr5	25818041	25818538	1	0	0	1
chr5	104330061	104332393	1	0	0	1
chr6	29747650	29749580	1	1	0	0
chr6	114340715	114342886	1	0	0	1
chr6	134931932	134932029	1	1	0	0
chr6	137002744	137003442	1	0	0	1
chr7	18500260	18500357	1	0	0	1
chr8	82218950	82219182	1	0	0	1
chr9	75081229	75084859	1	0	1	0
chr10	11352664	11353218	1	0	0	1
chr10	43560887	43564347	1	1	0	0
chr11	35729207	35733213	1	1	0	0
chr11	85673894	85674607	1	1	0	0
chr12	77032703	77032845	1	0	0	1
chr13	14557608	14557946	1	0	0	1
chr13	35038338	35039022	1	0	0	1
chr16	30717225	30719645	1	0	0	1
chr17	6936132	6936403	1	0	0	1
chr17	13550990	13553240	1	1	0	0
chr19	22588081	22588201	1	0	0	1
chrX	73848238	73848368	1	0	0	1
chrX	77676271	77680621	1	1	0	0
chrX	170734796	170737166	1	0	0	1
chrX	170809431	170816772	1	0	0	1
chrX	170819601	170820143	1	0	0	1
chrX	170841127	170846136	1	0	0	1
chrX	170853339	170853487	1	0	0	1
chr11	58628615	58628832	0	1	0	1
chr12	108517848	108517952	0	1	0	1
chr4	118548501	118548710	1	1	0	0
chr8	14306547	14307322	1	0	0	1
chr13	119598136	119600590	1	0	0	1

ATRi^{25}

ATRi^{55}

ATRi^{176} (- ATRi^{55} and ATRi^{25})

chrom	chromStart	chromEnd	ATRi+aph^{18hrs}	ATRi+aph^{9hrs}	ATRi^{18hrs}	ATRi+shTIM1
chr1	4258071	4258762	1	0	0	0
chr1	12657435	12657853	1	0	0	0
chr1	21220756	21220855	1	0	0	0
chr1	90444529	90445706	1	0	0	0
chr1	123303150	123304645	1	0	0	0
chr1	154180039	154180147	1	0	0	0
chr1	181742860	181743939	1	0	0	0
chr2	11717274	11717376	1	0	0	0
chr2	52064700	52069220	1	0	0	0
chr2	122375449	122375692	1	0	0	0
chr2	122377811	122378131	1	0	0	0
chr2	153856200	153858725	1	0	0	0
chr2	164704666	164705833	1	0	0	0
chr3	9801688	9804481	1	0	0	0
chr3	58253150	58253595	1	0	0	0
chr3	60045124	60046070	1	0	0	0
chr3	60736084	60736296	1	0	0	0
chr3	64170258	64172840	1	0	0	0
chr3	108615859	108616222	1	0	0	0
chr3	154608330	154611171	1	0	0	0
chr3	157440739	157441608	1	0	0	0
chr4	10137028	10138141	1	0	0	0
chr4	70884318	70884474	1	0	0	0
chr4	89264354	89265189	1	0	0	0
chr4	102493930	102495258	1	0	0	0
chr4	151738305	151738732	1	0	0	0
chr5	115809053	115809495	1	0	0	0
chr6	125414800	125419184	1	0	0	0
chr6	140114435	140114557	1	0	0	0
chr7	79466280	79467157	1	0	0	0
chr7	99410383	99412916	1	0	0	0
chr7	109844910	109847480	1	0	0	0
chr7	120835161	120835853	1	0	0	0
chr7	128187623	128188504	1	0	0	0
chr7	136870509	136870830	1	0	0	0
chr8	23277911	23278032	1	0	0	0
chr8	39152855	39153266	1	0	0	0
chr8	39155293	39155838	1	0	0	0
chr8	56972702	56972825	1	0	0	0
chr8	57571940	57572040	1	0	0	0
chr8	83756849	83756956	1	0	0	0
chr8	87092009	87092122	1	0	0	0
chr8	116557885	116559398	1	0	0	0
chr8	116562108	116563057	1	0	0	0
chr9	29483852	29484678	1	0	0	0
chr9	46132670	46133098	1	0	0	0
chr9	77720901	77721047	1	0	0	0
chr9	121927770	121929548	1	0	0	0
chr9	121931872	121933091	1	0	0	0
chr10	4055482	4055583	1	0	0	0
chr10	40645639	40647123	1	0	0	0
chr10	106280926	106281048	1	0	0	0
chr10	130594845	130594967	1	0	0	0
chr11	5169861	5170581	1	0	0	0
chr12	8955815	8955931	1	0	0	0
chr12	52929239	52930136	1	0	0	0
chr12	59120955	59121070	1	0	0	0
chr12	78350931	78351141	1	0	0	0
chr12	85733252	85733366	1	0	0	0
chr12	105196248	105196346	1	0	0	0
chr12	113424074	113425644	1	0	0	0
chr13	21169680	21170284	1	0	0	0
chr13	23400756	23401235	1	0	0	0
chr13	23517138	23517405	1	0	0	0
chr13	58349622	58350305	1	0	0	0
chr13	99775264	99776382	1	0	0	0
chr14	26687131	26688805	1	0	0	0
chr15	6526864	6529696	1	0	0	0
chr15	9015624	9015776	1	0	0	0
chr15	31714193	31714288	1	0	0	0
chr15	36908247	36909796	1	0	0	0
chr15	38975091	38975368	1	0	0	0
chr15	59178309	59182135	1	0	0	0
chr15	79461770	79463323	1	0	0	0
chr15	81215829	81217933	1	0	0	0
chr16	15670876	15673067	1	0	0	0
chr16	45352944	45353173	1	0	0	0
chr16	45552072	45552192	1	0	0	0
chr17	6932564	6932848	1	0	0	0
chr17	74084268	74087199	1	0	0	0
chr17	83170179	83170286	1	0	0	0
chr18	13296483	13299681	1	0	0	0
chr18	36801686	36802049	1	0	0	0
chr18	38107354	38107576	1	0	0	0
chr18	55371717	55371864	1	0	0	0
chr18	61320471	61320574	1	0	0	0
chr18	84831320	84831824	1	0	0	0
chr19	6813620	6815348	1	0	0	0
chr19	24912717	24912914	1	0	0	0
chr19	24918228	24918423	1	0	0	0
chr19	56847680	56848234	1	0	0	0
chrX	94883678	94883787	1	0	0	0
chrX	94969310	94969512	1	0	0	0
chrX	94979522	94979665	1	0	0	0
chrX	135040270	135041064	1	0	0	0
chrX	156064418	156064573	1	0	0	0
chrX	170830603	170830720	1	0	0	0
chr6	49236482	49236567	0	0	1	0
chr16	57391456	57391635	0	0	1	0
chr5	101209994	101210419	0	0	0	1

Figure 6. Tiered categorization of RPL peaks. List of RPL peaks that were detected in all 4 ATRi conditions (ATRi^{25}), in 2-3 ATRi conditions (ATRi^{55}), or in only 1 ATRi condition (ATRi^{176} (- ATRi^{55} and ATRi^{25})).

Genomic sites of ATR enrichment were compared to common genomic landmarks: chromatin state (euchromatic, heterochromatic, and boundary elements), gene location (transcription start sites and gene bodies) and common sequence elements (transcription factor binding sites and repetitive sequences). Of the 173 RPLs identified, ~22% were observed in euchromatin based on DNAse-hypersensitivity, early replication timing and euchromatin marks ($H3K4^{me3}$ and $H3K27^{ac}$). The remaining 78% of sites were observed in heterochromatin as assessed by a similar approach (DNAse-insensitivity and late replication timing). Nearly all the sites (112/173) were found outside of gene bodies, promoters and terminators, roughly reflecting the amount of non-coding DNA in the mammalian genome (Table 1). Notably, however, 21 of the 173 sites overlapped perfectly with CTCF binding sites (Table 1, 2, Figure 7) (Martin et al., 2011), which could not be explained by random coincidence given the low level of genomic coverage of both RPLs and CTCF binding sites (p-value = 0, Fisher's Exact Test). Nevertheless, the vast majority of CTCF sites (>99.9%) did not accumulate RPA following ATRi treatment, indicating that CTCF binding is not sufficient to cause fork collapse. Instead, the specific simple repeat sequence itself may influence fork collapse at the identified CTCF binding sites.

	Genic		Intergenic	Replication Timing Transition Regions (TTRs)	CTCF binding sites
	Intron	Exon			
Percentage of RPLs	**35.7%**		**64.3%**	**41%**	**12.3%**
	35.1%	**4.1%**			
Percentage in mouse genome	**41.1%***		**58.9%***	**32%***	**1.1%**
	37.7%*	**3.4%***			

***(Yue et al., 2014)**

Table 1. Genomic features of RPLs. Table describes the percent of RPL peaks that overlap with genic or intergenic regions, replication timing transition regions, and CTCF

binding sites. The bottom row provides comparison to the observed representation of each feature in the mouse genome.

RPL peak that overlaps with CTCF binding site (mm10)	Simple repeat associated with RPL
chr3:60045124-60046070	CACAG/CTGTG
chr5:104330061-104332393	CACAG/CTGTG
chr6:137002744-137003442	CACAG/CTGTG
chr6:29747650-29749580	CACAG/CTGTG
chr10:11352664-11353218	CACAG/CTGTG
chr13:74528418-74531130	CACAG/CTGTG
chr15:38975091-38975368	CACAG/CTGTG
chr17:13550990-13553240	CACAG/CTGTG
chr17:13305944-13308113	CACAG/CTGTG
chr5:25818041-25818538	CACAG/CTGTG, CACACACAG/CTGTGTGTG
chr13:35038338-35039022	CACAG/CTGTG, CACACAG/CTGTGTG
chr6:86329935-86334080	CAGAGG/CCTCTG
chr6:90513949-90522651	CAGAGG/CCTCTG
chr6:125414800-125419184	AGGCAGG/CCTGCCT
chr9:121927770-121929548	CAGG/CCTG
chr11:5169861-5170581	CGGTGCCTGACATACAC/GTGTATGTCAGGCACCG
chr16:45352944-45353173	ACAGACAGG/CCTGTCTGT
chr19:6813620-6815348	TCACCATGCAGGACTTG/CAAGTCCTGCATGGTGA
chr2:122377811-122378131	-
chr4:118221107-118221315	-
chr7:120835161-120835853	-

Table 2. Comprehensive list of RPL peaks and associated simple repeats that overlap with CTCF binding sites.

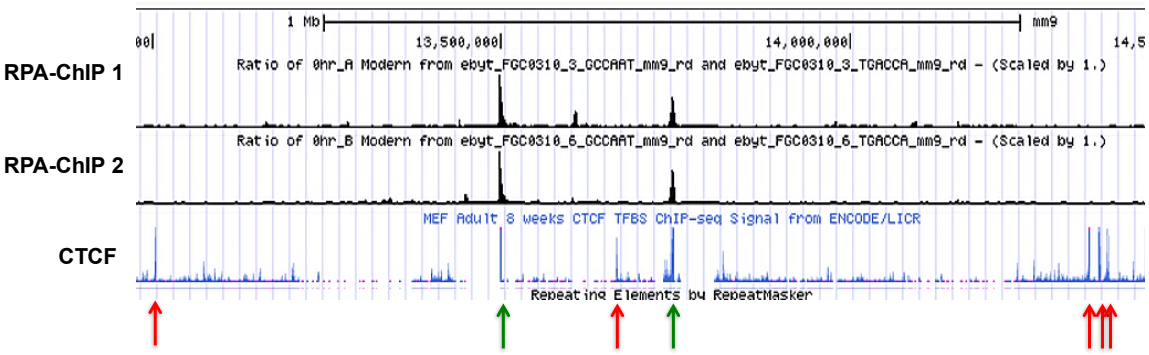

Figure 7. Overlap of RPL peaks with CTCF binding sites. Coverage tracks from two biological replicates of ATRi+aph^{18hrs} are depicted with the CTCF ChIP-Seq track from UCSC genome browser. Green arrows denote overlap between CACAG/CTGTG-centered RPL peaks and CTCF peaks; red arrows denote CTCF peaks that do not correspond to observable RPL peaks.

A previous report of RPA accumulation sites, termed ERFS for Early Replicating Fragile Sites (Barlow et al., 2013), were also examined for overlap with RPLs. ERFS hotspots are reportedly composed of broad, low-amplitude peak regions that mark clusters of co-localized RPA, BRCA1 and SMC5 binding and cover approximately 4.7% of the mouse genome in aggregate (Barlow et al., 2013). A total of 20 out of the 173 identified RPLs (~12%) were found within ERFS (p-value = 0.0001, Fisher's Exact Test), and, although significant, the overlap of these sites likely reflects the breadth of ERFS across the mouse genome. Moreover, because ERFS were identified in cells synchronized at the G1/S transition, and thus predominantly mark sites that are adjacent to origins of DNA replication, considerable overlap was not expected since our RPA-ChIP procedures sampled the entire mouse genome. Notably, common fragile sites (CFS), megabase-sized sites of chromatid breaks in M phase, were also not disproportionally enriched in RPLs. This observation is consistent with recent studies that DNA replication through CFS is indistinguishable from replication through non-fragile regions and that chromatid breakage correlates best with a low density of origin firing and continued replication into M phase (Letessier et al., 2011; Minocherhomji et al., 2015; Ghamrasni et al., 2016).

Despite the lack of correlation with previously characterized chromatin features and breakage sites, close inspection of raw read enrichment at RPLs revealed two key characteristics: 1) a seemingly high frequency of various repetitive sequences and elements (LINE, SINE, LTRs), and 2) a distinct absence of aligned reads at the approximate center of these peaks, which according to the reference genome, contains simple tandem repeats of lengths greater than 100 bps (Figure 8). The absence of reads aligned to long lengths of tandem repeats is expected as reads comprised entirely of these

repetitive sequences should map to multiple genomic locations and thus be binned by filtering programs. The position of the simple tandem repeats at the center of the RPL peaks, as well as the apparent abundance of LINE, SINE and LTR elements within RPL peaks, led to the hypothesis that such features may contribute to replication fork collapse at these sites following ATR inhibition.

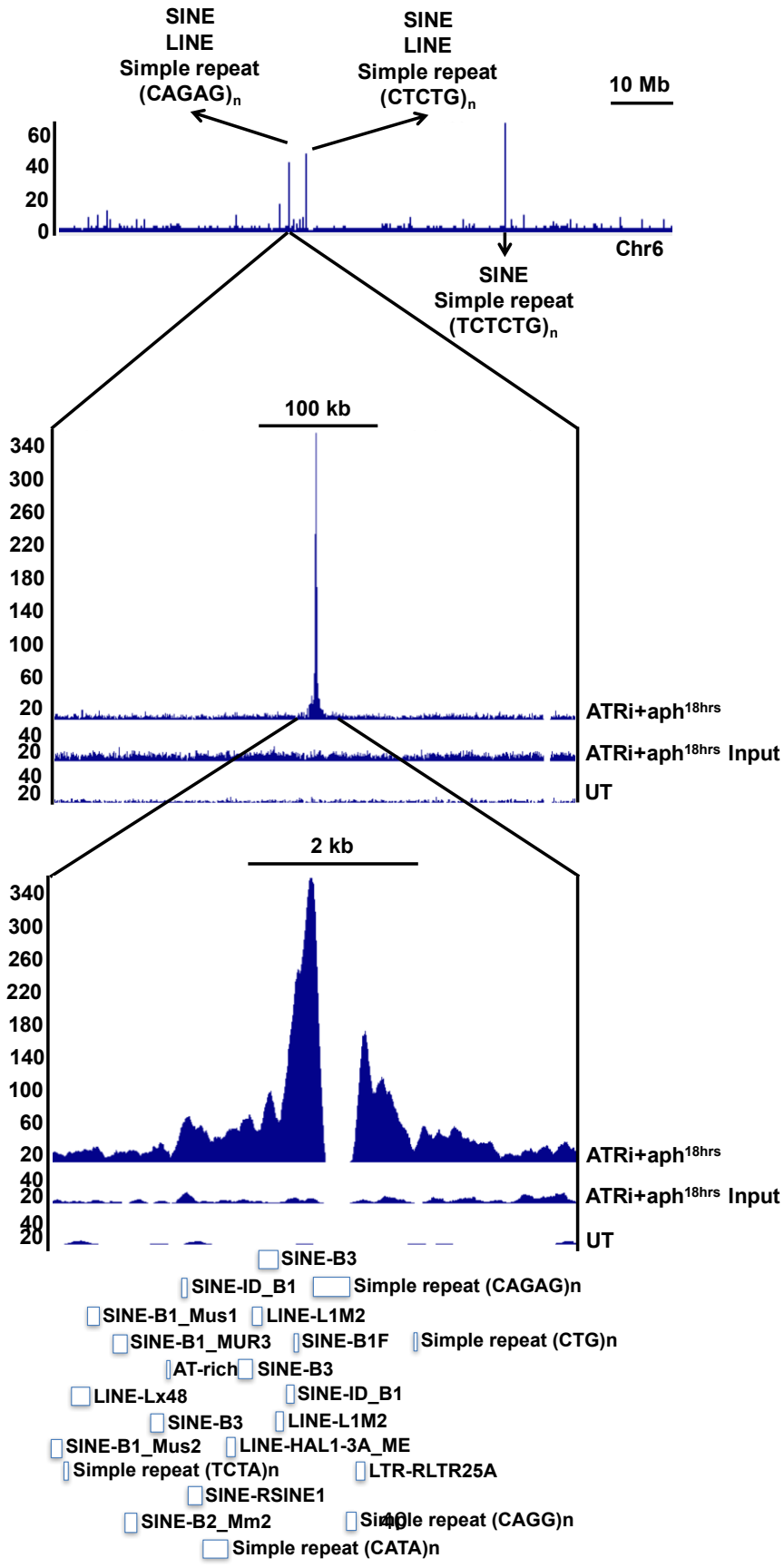
SINE
LINE
Simple repeat
(CAGAG)n
SINE
LINE
Simple repeat
(CTCTG)n
10 Mb
60
40
20
0
Chr6
SINE
Simple repeat
(TCTCTG)n
100 kb
340
300
260
220
180
140
100
60
20
40
20
40
20
ATRi+aph18hrs
ATRi+aph18hrs Input
UT
2 kb
340
300
260
220
180
140
100
60
20
40
20
40
20
ATRi+aph18hrs
ATRi+aph18hrs Input
UT
SINE-B3
SINE-ID_B1
Simple repeat (CAGAG)n
SINE-B1_Mus1
LINE-L1M2
SINE-B1_MUR3
SINE-B1F
Simple repeat (CTG)n
AT-rich
SINE-B3
LINE-Lx48
SINE-ID_B1
SINE-B3
LINE-L1M2
SINE-B1_Mus2
LINE-HAL1-3A_ME
Simple repeat (TCTA)n
LTR-RLTR25A
SINE-RSINE1
SINE-B2_Mm2
Simple repeat (CAGG)n
Simple repeat (CATA)n

Figure 8. RPA peaks along part of chromosome 6 of the mouse genome. Top track depicts the ratio of RPA ChIP over input for ATRi+aph^{18hrs}. Arrows detail the repetitive elements present in each RPA peak. A single RPA peak is zoomed in to reveal repetitive regions within the site of RPA accumulation. Top row depicts the coverage track of ATRi+aph^{18hrs} (RPA ChIP); the second row depicts the coverage track of its input (RPA Input); and the third row depicts the coverage track of UT (RPA ChIP). RepeatMasker annotations of repetitive elements in the specified region are depicted at the bottom. The gap in read accumulation observed in the coverage tracks overlaps with $(CAGAG)_n$ repeats.

In aggregate, of the 173 RPL peaks identified in ATRi+aph^{18h}, 71 were centrally localized around CAGAGG/CCTCTG simple repeats, 27 were centered around CACAG/CTGTG repeat sequences, and 3 contained CAGG/CCTG tandem repeats (Table 3). However, quantification of repeat enrichment in ChIP samples through sequences derived from the reference genome is not entirely accurate, as the mapping of highly repetitive regions remains difficult by current standards of reference genome assembly. Reference assemblies are typically updated and released with new versions every few years as better methods of covering repetitive regions, such as through long-read technology, are applied, indicating the inherent challenges of correctly representing long stretches of repetitive sequences.

Simple Repeat	Number of RPL Peaks with Repeat	Length of Repeat (monomer units)	Number of peaks in ATRi^{55}	Number of peaks in ATRi^{25}
CAGG/CCTG	3	11 - 618	0	0
CACAG/CTGTG	27	17 - 413	15	1
CAGAGG/CCTCTG	71	9 - 184	22	22
CAGAGT/ACTCTG	1	42	0	1

Table 3. Summary of repeats found enriched in the ATRi+aph^{18hrs} RPA ChIP. The second column notes the number of RPL peaks that contained the designated simple repeat. The third column denotes the minimum and maximum range of the repeat length within the relevant RPL peaks. The fourth and fifth columns denote the number of the repeat-containing peaks that are a part of the tiered categorization of ATRi-sensitive RPL peaks.

2.2 REQer: Enrichment of repetitive sequences in RPLs

To evaluate the association of repeats and elements observed within RPL peaks to the level of RPA enrichment without relying on the reference genome sequence, the number of reads from retrieved RPA ChIP-Seq samples that contained the observed elements and tandem repeats were counted. In addition, other repeat-containing genomic elements and regions, such as ribosomal DNA, tRNA genes, and satellite DNA were also analyzed. As expected from previous studies in yeast (Brewer and Fangman, 1988; Gruber M et al., 2000; Weitao T et al., 2003), ribosomal DNA was enriched by two-fold in ATRi+aph^{18hrs} treated cells compared to vehicle-treated controls (Figure 9). However, LINE, SINE, and LTR elements, as well as other common genomic repeats, such as tRNA genes and satellite repeats, were not increased in ChIP retrievals from cells treated with ATRi and low dose aphidicolin (Figure 9). These data indicate that the apparent accumulation of LINE, SINE and LTR elements in peak regions were not significantly enhanced over their genomic representation. Thus, these elements do not appear to be sufficient to cause replication fork collapse upon ATR inhibition.

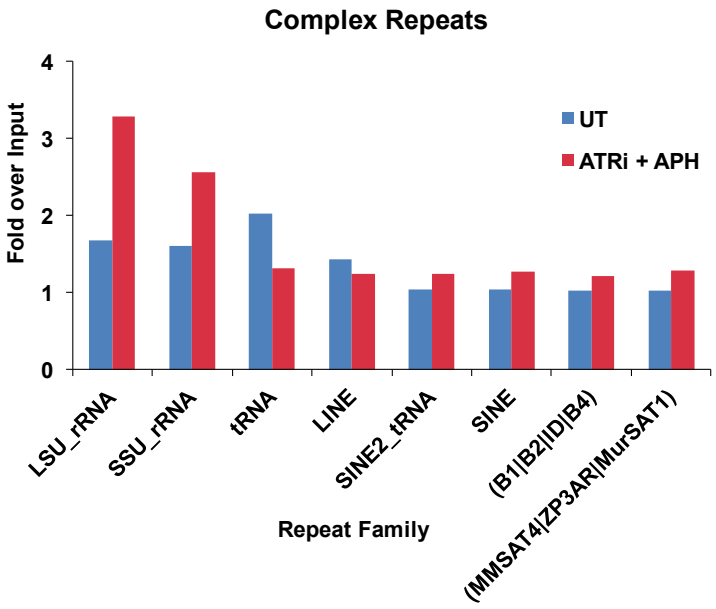

Figure 9. Complex repeat enrichment in RPA ChIP samples. Bar graph depicts the fold enrichment values of different families of complex repeats counted and averaged among replicates from each condition: ATRi+aph^{18hrs} (ATRi + APH) and untreated (UT). These values were normalized by the total number of reads with at least one reported alignment.

Simple tandem repeats found at peak centers or within RPA-accumulation regions were also queried for enrichment in RPA-ChIP retrievals. However, accurate quantification of these repeats within reads required the development of new repeat counting programs, since southern blot detection of two independent RPLs indicated that repeats within these sites were polymorphic in length across different mouse strains and some simple tandem repeats contain frequent repeat interruptions (data not shown).

Therefore, we generated a program that accurately counts the number of repeat monomers within reads both in tandem and in aggregate (REQer, Repeat Enrichment Quantifier). As monomer count within a read increases, a larger fraction of the read is comprised of monomers until the maximum length of the read is reached. The presence of various simple repeats within reads from RPA-ChIP retrievals normalized by the counts

observed in respective input reads was directly quantified, bypassing the need for alignment to the reference genome. Sequenced reads from each sample, after adapter trimming, were just below 100 bp in length. Duplicate reads were removed and three biological replicates were combined for each condition: ATRi+aph^{18hrs} RPA-ChIP, ATRi+aph^{18hrs} Input, Untreated (DMSO) RPA-ChIP, and Untreated (DMSO) Input.

Repeats found within RPL peaks, as well as those that previous studies had shown to be difficult to replicate, were tested for enrichment by REQer. In the first analysis, the program measured the occurrences of increasing tandem lengths of each simple repeat and its complement within all sequencing reads of each condition and their respective inputs. The graphed quantifications are displayed in Figure 10.

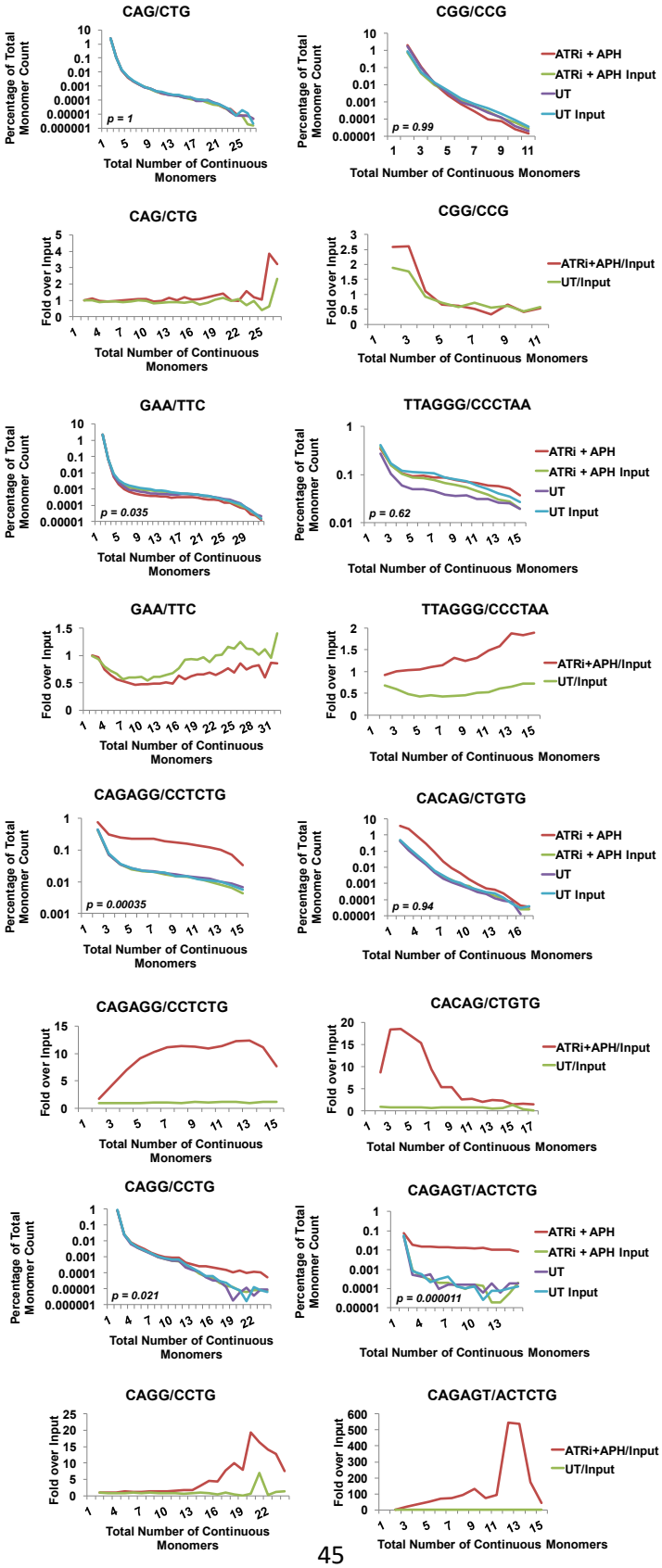
CAG/CTG
Percentage of Total Monomer Count
p = 1
Total Number of Continuous Monomers
CGG/CCG
p = 0.99
ATRi + APH
ATRi + APH Input
UT
UT Input
Fold over Input
ATRi+APH/Input
UT/Input
GAA/TTC
p = 0.035
TTAGGG/CCCTAA
p = 0.62
CAGAGG/CCTCTG
p = 0.00035
CACAG/CTGTG
p = 0.94
CAGG/CCTG
p = 0.021
CAGAGT/ACTCTG
p = 0.000011

Figure 10. Tandem simple repeat analysis of RPA ChIP-Seq samples. Analysis was performed on sequencing reads from combined biological replicates of each condition: ATRi+aph^{18hrs} (ATRi + APH) and untreated (UT). The occurrences of specified tandem monomers of each repeat within all sequencing reads of combined replicates were measured and graphed. The x-axis depicts the range of tandem units of the repeat queried and the y-axis depicts the frequency of the occurrences as a percentage of total monomer count of the repeat present in the reads of the combined replicates (top row). At each specified number of tandem monomers, the ratio of the ATRi + APH value over its input and of the UT value over its input was graphed to depict fold over input enrichment (bottom row). P-values were obtained at 95% confidence interval using the Kolmogorov-Smirnov test between the distributions of ATRi+aph^{18hrs} and its input and are indicated.

The x-axis displays the number of tandem units of the repeat, with the highest range signifying the maximum number of times a repeat of particular length can occur within a 90 bp sequence read. The y-axis displays the frequency of each tandem monomer length as the percentage of total monomer count of the repeat sequence in the combined replicates. Notably, triplet repeats, which have previously been shown to impede DNA replication and cause an increased reliance on ATR orthologues for stability when expanded, were only mildly enriched in ATRi+aph^{18hrs} RPA-ChIP reads, consistent with their wild-type lengths (Figure 10). These include common triplet sequences implicated in various disease-states (e.g. expanded CGG/CCG repeats in fragile X syndrome, expanded CAG/CTG repeats in Huntington's disease, and expanded GAA/TTC repeats in Friedreich's ataxia), all of which did not result in significant differences between the ChIP retrieval and input, indicating little effect of these repeats on fork collapse (Figure 10). Telomere repeats (TTAGGG), which have also been demonstrated to be difficult to replicate and require ATR for fork stability (Sfeir et al., 2009), were enriched approximately 2-fold by the ATRi+aph^{18hrs} condition (Figure 10). Notably, simple tandem repeats that demonstrated the greatest levels of enrichment were made up of monomers that had not been previously noted as difficult to replicate (Figure 10). Enrichment levels for each

simple repeat are summarized in Figure 10. These repeat monomers included hexameric and pentameric repeats: CAGAGG/CCTCTG, CAGAGT/ACTCTG, and CACAG/CTGTG (Figure 11). The shortest monomer that was significantly enriched in RPA ChIP-Seq retrievals, CAGG/CCTG, is the same repeat known to cause human myotonic dystrophy type 2 (Liquori et al., 2001). However, the two RPL peaks that harbored a little over 300 and 600 tandem runs of the repeat in the mouse genome were not in the murine myotonic dystrophy type 2-associated gene, but instead were found in non-genic regions.

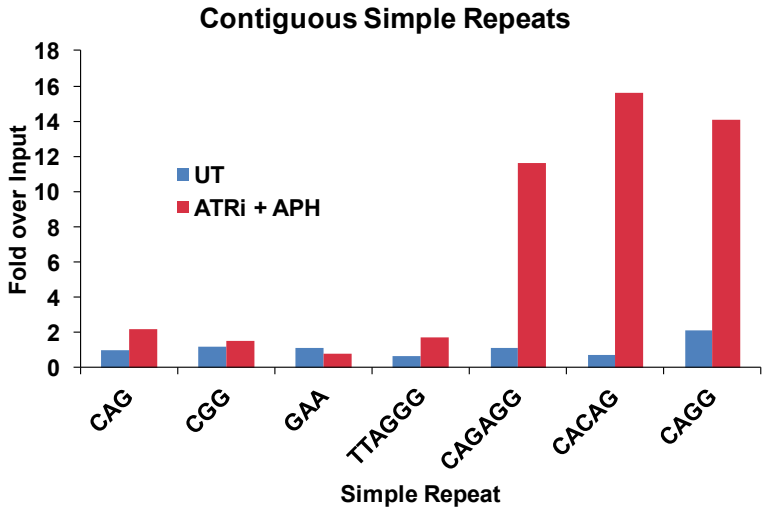

Figure 11. Tandem simple repeat enrichment in RPA ChIP samples. Bar graph depicts the average fold-enrichment values at the highest level of difference between the ChIP retrieval and input for ATRi+aph^{18hrs} (ATRi + APH) and untreated (UT) conditions.

REQer was next used to quantify simple repeat enrichment without necessitating tandem occurrence of the repeats. Consideration of simple repeats that permit slight interruptions of intervening sequence would address the importance of repeats that only occur continuously to cause fork collapse. The reference genome sequence at certain

RPL sites suggests that partial interruption within repetitive sequences is occurring. Thus, by a second analysis through REQer, focus was placed on non-contiguous repeat enrichment. Reads were categorized by the aggregate monomer count of each simple repeat within a read. Their frequency was graphed as the fraction of reads within each condition that contained the specified amounts of repeat monomers (Figure 12). By this method of analysis, the distribution of CACAG/CTGTG repeats in RPA ChIP-Seq retrievals is significant (p-value < 0.005, Kolmogorov-Smirnov test), indicating that the increased occurrence of CACAG/CTGTG sequences within a distinct region, even if it is not tandem, is still sufficient to lead to RPA accumulation and fork collapse.

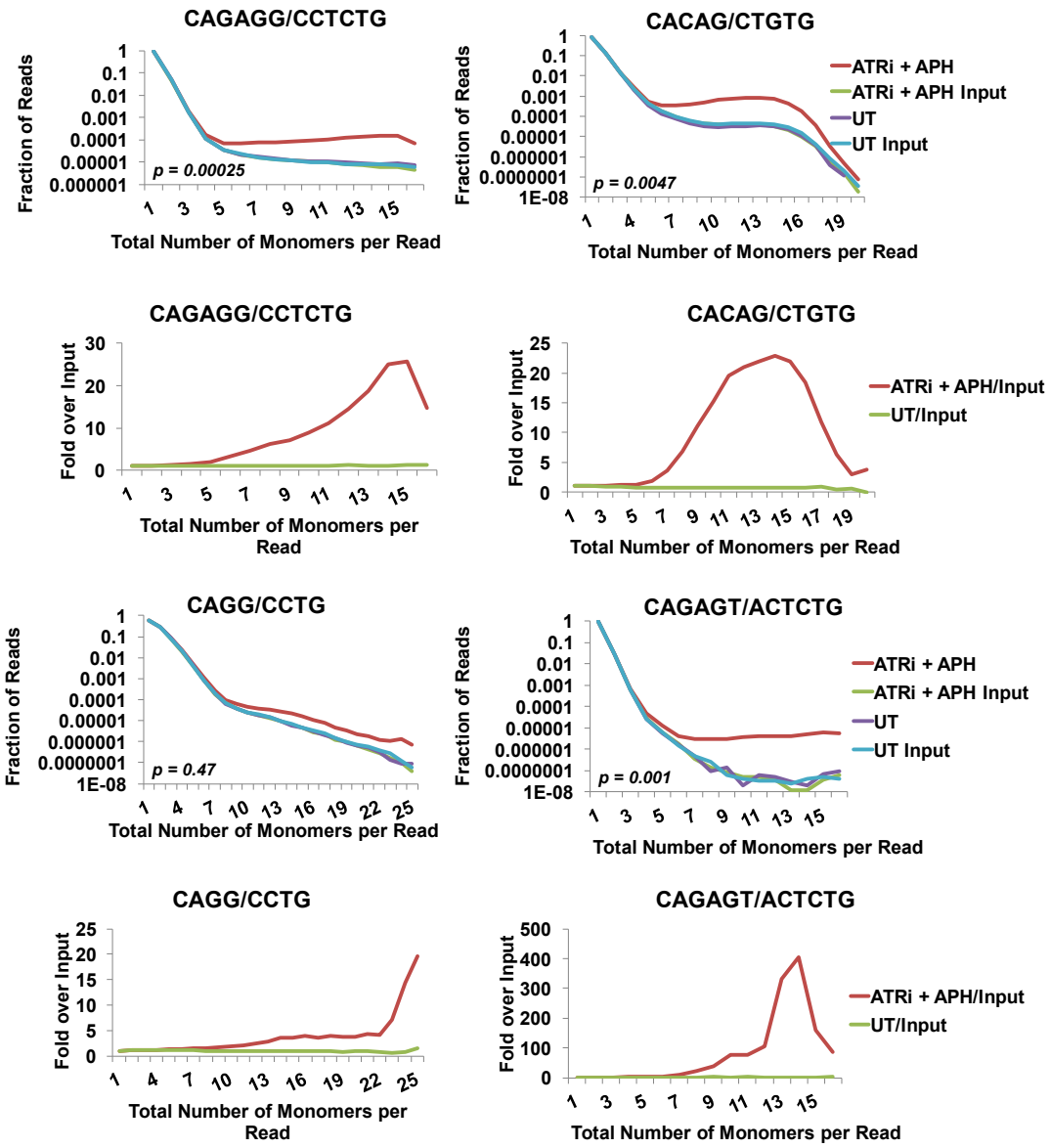

Figure 12. Non-contiguous simple repeat analysis of RPA ChIP-Seq samples. Analysis was performed on sequencing reads from combined biological replicates of each condition: ATRi+aph^{18hrs} (ATRi + APH) and untreated (UT). Sequencing reads were categorized by the aggregate monomer count of each repeat within a read, allowing for intervening sequence between repeats. Their frequency was calculated as the fraction of reads within the total number of reads from combined replicates that contained the specified amounts of repeat monomers (top row). At each specified number of total monomers per read, the ratio of the ATRi+aph^{18hrs} value over its input and of the UT value over its input was graphed to depict fold over input enrichment (bottom row). P-values were obtained at 95% confidence interval using the Kolmogorov-Smirnov test between the distributions of ATRi+aph^{18hrs} and its input and are indicated.

Notably, some of the most highly enriched repeats in ATRi+aph^{18hrs} RPA-ChIP-Seq retrievals according to REQer (Figures 10, 11 and 12), such as CAGAGG/CCTCTG and CACAG/CTGTG (hereon abbreviated as CAGAGG and CACAG repeats), were also observed in the largest number of RPL peaks (Table 3). To determine if the number of repeats correlates with replication fork collapse upon ATRi treatment, we assessed the lengths of CACAG and CAGAGG repeats across the mm10 reference genome. This analysis demonstrated a significantly higher monomer count of these repeats in RPLs than what was observed across the genome, thus indicating that greater repeat lengths increase the incidence of fork collapse following ATRi treatment (Figure 13).

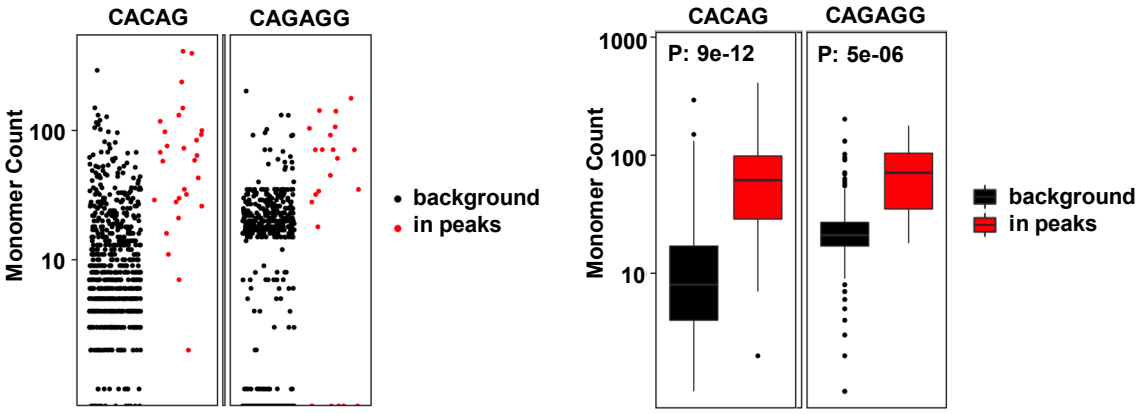

Figure 13. Comparison of repeat length in RPLs to repeat length in the genome. Box-plot depicts representation of CACAG and CAGAGG repeats in the mouse genome compared to representation of the repeats in RPL peaks from ATRi+aph^{18hrs}-treated cells. Black dots denote each occurrence of the repeat in the genome at the observed monomer count; red dots denote each occurrence of the repeat in the RPL peaks at the observed monomer count. This analysis utilized the RepeatMasker database's annotation of known repeats to intersect their locations with RPL peaks. The amount of intersection was compared to randomly shuffled peaks. P-values were calculated by Wilcoxon signed–rank test and are indicated.

Of note, some repeats found enriched in RPA ChIP-Seq retrievals from ATRi-treated cells were imbedded within extensive stretches of repeats of similar but distinct

sequences (e.g. CAGAGT/ACTCTG occurring adjacent to CAGAGG/CCACAG repeats). These repeat orthologues were only found in a fraction of RPLs that contained more frequently observed repeats, and thus were not essential for fork collapse. These associations suggest that rare variant repeats, such as CAGAGT, might be “passenger repeats” that become enriched during retrieval due to fork collapse from difficulties in replicating other more problematic and extensive repeats, such as CAGAGG. Nevertheless, these repeat concurrences made linking replication fork collapse to any specific repeat more ambiguous and implied the need for additional methods to examine the role of each repeat in fork collapse.

Overall, analysis of raw sequencing reads from RPA ChIP-Seq samples identified lengthier simple repeats as novel regions of replication fork collapse in cells experiencing replication stress from ATRi and aphidicolin treatment. It has been previously speculated that replication fork slowing occurs preferentially at common fragile sites and expanded triplet repeat sequences. These data, however, suggest that a new and distinct class of sequences is prone to fork collapse, namely pentameric and hexameric simple repeats that occur in tandem or with partial interruption, indicating that the inherently complicated features of longer repeats may drastically compromise smooth progression of the replication fork through the region.

2.3 Simple tandem repeats in RPLs form stable intrastrand structures

The enrichment of RPA within RPLs is consistent with the normal means of ATR activation and reliance on it for fork stability (Shechter et al., 2004; MacDougall et al., 2007; Flynn and Zou, 2011). Generation of ssDNA can be the product of impeded polymerase progression, elevated nascent strand resection at forks with intact parental

strands, or resection of DSBs at replication forks after collapse (Zegerman and Diffley, 2009). The need for ATR at these sites to prevent site-specific accumulation of RPA indicated that some defect in DNA replication might precede and promote ATR activation.

The Polα, Polδ and Polε replicative polymerase complexes can be impeded by numerous abnormalities, including abasic sites, damaged bases, and the formation of intrastrand secondary structures (Aguilera and García-Muse, 2013). However, the prominent enrichment of simple tandem repeats at ATRi-sensitive sites suggested the possibility that these sequences could either be more vulnerable to damage or more prone to form intrastrand secondary structures when disassociated from their complementary strands. Repeat-dense regions, particularly upon unwinding of DNA duplexes, may physically impede progressing polymerases at the replication fork by forming secondary structures. Because of the numerous examples of simple tandem repeats forming secondary structures, we examined the structure-forming potential of synthetic oligonucleotides (oligos) of the repeat sequences identified within RPLs.

Synthetic single-stranded oligos of RPL repeats found to be enriched in RPA-ChIP retrieval (Table 4) were examined by native polyacrylamide gel electrophoresis (PAGE) after one annealing cycle of heating and re-cooling to room temperature (Figure 14). A single dominant band was observed for most of the oligos, and each of these demonstrated greater mobility than expected based on their length (Figure 14), which is an indicator of the formation of a compact intrastrand secondary structure. The presence of well-defined bands suggests that the structures are uniquely folded. $(CAGAGT)_{15}$ was the only sequence that showed a significantly slow-moving band, which could be attributed to loosely bound dimer. Notably, three purine-rich oligos, CAGAGG, CAGAGT, and CAGAGAGG, exhibited greater electrophoretic mobility than oligos encoding their

complementary strands (Figure 14), suggesting that the purine-rich strands of these repeats have more structure-forming potential than their pyrimidine-rich complements.

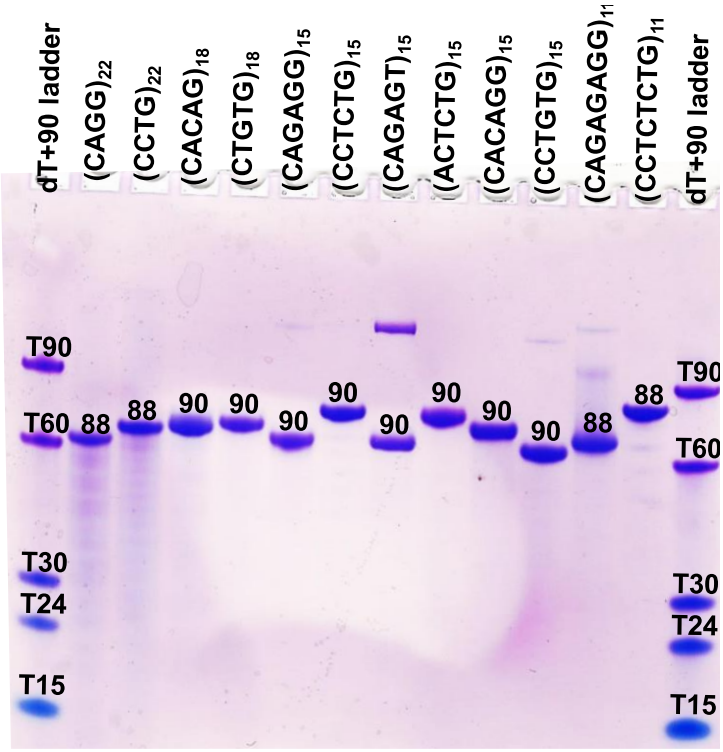

Figure 14. Non-denaturing 12% PAGE gel of repeats listed in Table 4. The gel was run in 1x TBE supplemented with 3 mM $MgCl_2$ for 160 minutes at 150 V. The DNA bands were visualized with Stains All. Samples were prepared in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

To more conclusively determine the ability of these repeats to form stable secondary structures, the specific folding of the repeat-containing oligos was examined by circular dichroism (CD) scans and thermal difference spectra (TDS), which quantify base interactions, uniformity of structure, and enthalpic stability (Table 4).

Sequence	nt	TDS (nm)	CD: molar ellipticity, $M^{-1}cm^{-1}$ (λmax, nm)	T_m, °C Heating	ΔH, kJ/mol heating
5'- $(CAGG)_{22}$ -3'	88	269 256	-324 ± 20 (280.6) 182 ± 52 (261)	NF	-
5'- $(CCTG)_{22}$ - 3'	88	277 241	375.2 ± 5.3 (284) -317.4 ± 3.4 (254.6)	NF	-
5'- $(CACAG)_{18}$ -3'	90	265	575 ± 27 (289) 654 ± 34 (280) -543 ± 80 (253)	48.9 ± 0.5 hysteresis 3.8°C	100. ± 6
5'- $(CTGTG)_{18}$ -3'	90	276	90. ± 1.0 (291) 82 ± 24 (243) -296 ± 17 (264)	50.4 ± 1.2*	
5'- $(CAGAGT)_{15}$ -3'	90	262.5	48.9 ± 0.2 (280.3) -76.9 ± 3.7 (267) 97.3 ± 0.4 (249)	NF	-
5'- $(ACTCTG)_{15}$ -3'	90	269	317 ± 15.3 (280) -334 ± 40 (241.3)	NF	-
5'-$(CACAGG)_{15}$ -3'	90	266 259 (shoulder)	255 ± 4.2 (277.5) -318 ± 26 (248)	NF	-
5'-$(CCTGTG)_{15}$ -3'	90	278 236	158 ± 9 (285.3) 317 ± 8 (261.3)	51.5 ± 1.1 hysteresis 5.3°C	282.6 ± 13.4
5'-$(CAGAGAGG)_{11}$ -3'	88	260	728 ± 68 (262.5) 373 ± 12 (276)	46.4 ± 0.6 hysteresis 4.4°C	209.7 ± 6.6
5'-$(CCTCTCTG)_{11}$ -3'	88	280 235 (shoulder)	444 ± 22 (279) -251.5 ± 9.2 (251.5)	NF	-
5'-$(CAGAGG)_{15}$-3'	90		1097 ± 69 (261) -686 ±12 (241) shoulder (276) shoulder (291)	55.6 ± 0.9	223 ± 13
5'-$(CCTCTG)_{15}$ -3'	90	279	444 ± 28 (282.5) -200 ± 7 (253)	NF	-

*transition is not well defined. The values of ΔH and hysteresis cannot be accurately determined

Table 4. Biophysical parameters of RPL repeat sequences. TDS – thermal difference spectra; CD – circular dichroism; for both CD and TDS, maxima and minima on the curve are reported. Tm Fit – melting temperature obtained using fitting procedure that assumes temperature independent enthalpy (e.g. ΔH = const, heat capacity, Cp = 0). NF stands for no fit. Repeat number for tandem sequences was adjusted to maintain the constant length of 88-90 nt.

The stability of the repeat-containing oligos was determined in CD and UV-vis melting studies (Table 4). Thermal difference spectra were generated by subtracting UV-vis spectra taken at 90°C from those taken at 4°C. The TDS signature of pyrimidine-rich strands with their maxima at ~280 nm and shoulders at ~235 nm was consistent with DNA self-complementary duplex with 50-100% GC content (Mergny et al., 2005). The TDS

signature of purine-rich strands with their maxima at ~260 nm could not be assigned to one particular secondary structure.

Circular dichroism (CD) wavelength scans were also performed to detect the characteristic signatures of secondary structures formed by the sequences tested. The CD scan of all the oligos displayed varied and mostly unusual folding. The values of molar ellipticity at the wavelength for the most intense band was used to correlate the extent of DNA folding (Figure 15).

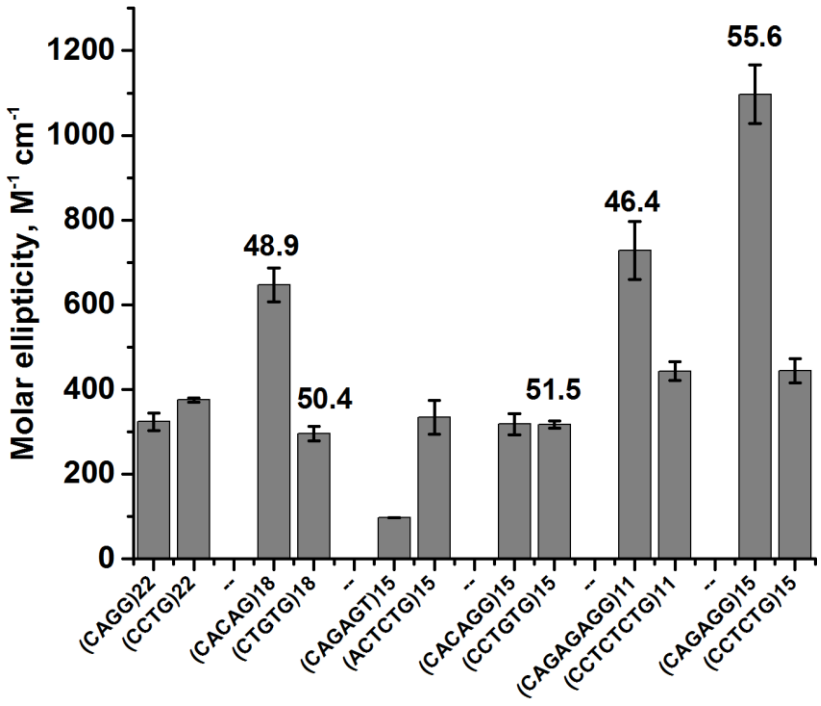

Figure 15. CD molar ellipticity of RPL repeat sequences. Absolute values of CD molar ellipticity for the most intense peak (different in each case) are shown as a proxy for the amount of structural folding of the DNA. Only 5 out of the 12 DNA sequences demonstrated clear melting transition. Melting temperatures obtained for those cases used non-linear fit assuming a two-state system and are shown above the relevant graph bar. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$). The data are represented as mean +/- SEM.

Oligos with the highest values of molar ellipticity were CACAG ($\Delta\varepsilon$ = 645 $M^{-1}cm^{-1}$), CAGAGG ($\Delta\varepsilon$ = 1097 $M^{-1}cm^{-1}$), and CAGAGAGG ($\Delta\varepsilon$ = 728 $M^{-1}cm^{-1}$), and were assumed to be the most folded. During melting, only a few sequences displayed clear transitions to allow determination of T_m (Figure 15, Table 5). The three purine-rich repeats with extensive secondary structure exhibited stability substantially higher than 37°C (CACAG: T_m = 49°C, ΔH = 100 kJ/mol; CAGAGG: T_m = 56°C, ΔH = 223 kJ/mol; CAGAGAGG: T_m = 46°C, ΔH = 209 kJ/mol). Their melting transition was reversible or nearly reversible (hysteresis was small, <5°C), implying that the structures formed were unimolecular (intrastranded). Importantly, with the exception of CACAGG, oligos encoding the complementary strand of these repeats neither appeared to have uniform structures based on their weak CD signature, nor displayed clear melting transition (Table 4 and Figure 15) and, as such, were probably mostly unstructured. While other repeats also exhibited some features of structure formation or stability (Table 4), our combined data suggested that CACAG, CAGAGG and CAGAGAGG demonstrated characteristics of unimolecular secondary structure formation more so than other repeats found in RPLs.

The CAGAGG repeat demonstrated the highest stability and uniformity in structure formation (Table 4 and Figure 15), and notably, it was also observed frequently in RPL peaks (Table 3). For these reasons, we further characterized the requirements for structure formation of this repeat sequence. By CD signature and thermal stability, we observed that no fewer than five tandem units of CAGAGG were required to generate stable secondary structure (Figures 16, 17, 18, 19 and Table 5). The CD signature with a major peak at ~260 nm and two prominent shoulders at 276 and 291 nm was unusual and did not correspond to any commonly known DNA secondary structure (Figure 16).

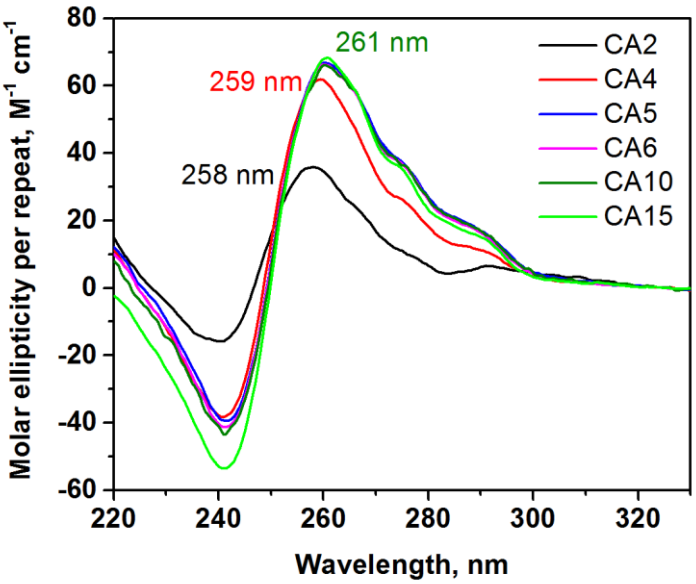

Figure 16. CD wavelength scans normalized per CAGAGG repeat. Scans show strength of CD signal as a proxy for the amount of structural folding of the DNA. CA2 = $(CAGAGG)_2$; CA4 = $(CAGAGG)_4$; CA5 = $(CAGAGG)_5$; CA6 = $(CAGAGG)_6$; CA10 = $(CAGAGG)_{10}$; CA15 = $(CAGAGG)_{15}$. Peak maxima are annotated above the curves. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

The superposition of CD signature for $(CAGAGG)_5$-$(CAGAGG)_{15}$ (Figure 16) suggested that each $(CAGAGG)_n$ oligo, when folded, consisted of the same basic structural units. Interestingly, the melting temperature did not increase significantly for oligos with n > 5, reaching only 56 ± 1°C for $(CAGAGG)_{15}$ (Figure 18, Table 5). This finding indicated that a minimum of five CAGAGG repeats were required to form the basic structural unit; oligos with a higher number of repeats most likely contain multiple structural units that do not interact significantly with each other ('bead-on-a-string model').

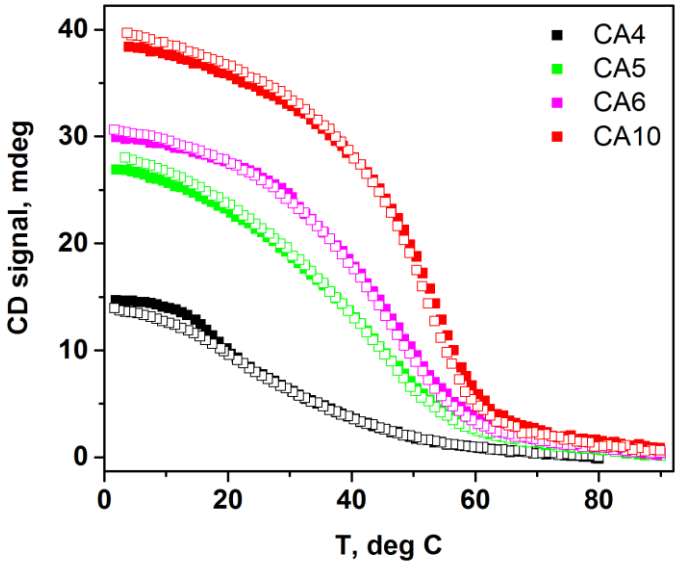

Figure 17. Representative CD melting curves for $(CAGAGG)_n$ (n = 4, 5, 6 and 10). Melting curves demonstrate the reversibility of melting transition. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

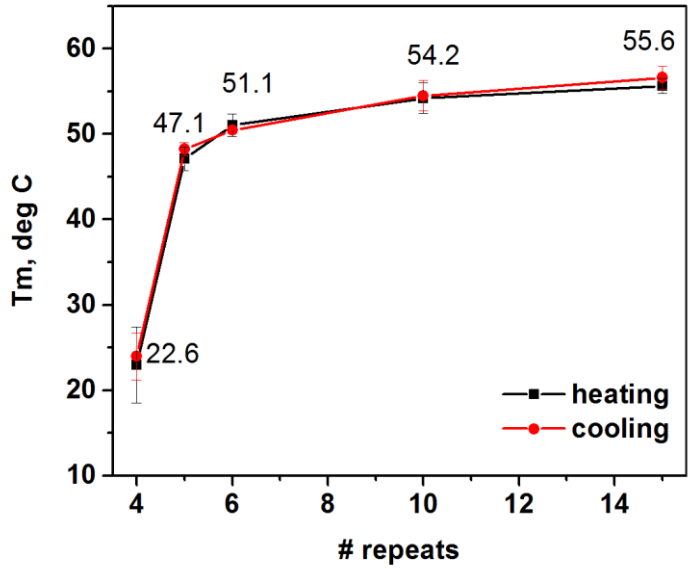

Figure 18. Graph of melting temperatures obtained in UV-vis melting studies at different monomer lengths of the CAGAGG repeat. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

Repeat	T_m, °C, heating	T_m, °C, cooling	ΔH, kJ/mol, heating
*CA4	23 ± 4	24 ± 3	98 ± 3
CA5	47.1 ± 1.4	48.3 ± 0.8	117 ± 9
CA6	51.1 ± 1.3	50.5 ± 0.4	95 ± 14
CA10	54.2 ± 1.8	54.5 ± 1.8	166 ± 14
CA15	55.6 ± 0.9	56.6 ± 1.4	222 ± 3

the structure is too unstable for accurate determination of T_m.

Table 5. UV-vis melting data on increasing lengths of CAGAGG repeat. T_m values, obtained by both heating and cooling, and change in enthalpy are summarized. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

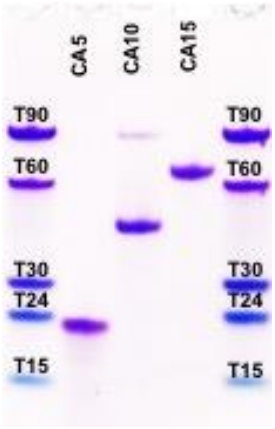

Figure 19. Non-denaturing 12% PAGE gel of CA5, CA10, and CA15. Samples were in 1xTAC buffer supplemented with 3 mM $MgCl_2$ run at 140 V for 115 minutes. The DNA bands were visualized with Stains All. CA5 = $(CAGAGG)_5$; CA10 = $(CAGAGG)_{10}$; CA15 = $(CAGAGG)_{15}$. Samples were prepared in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

To confirm the unimolecular nature of the secondary structure formed by CAGAGG, we measured CD spectral changes and thermal stability across a >20-fold change in concentration for $(CAGAGG)_5$ and $(CAGAGG)_{10}$ (Figures 20 and 21). Figure 20 displays overlay of normalized UV-vis melting data for $(CAGAGG)_{10}$ at varying strand

concentrations. The graph presented in Figure 21 depicts CD scans at 4°C of the samples after UV-vis melting. At all concentrations tested, the measured values were invariable (Figures 20 and 21), supporting a unimolecular nature of the structural unit. These results signify that CAGAGG tandem repeats observed in RPLs, averaging over 100 monomer units, can form secondary structures at physiological temperatures and ionic compositions.

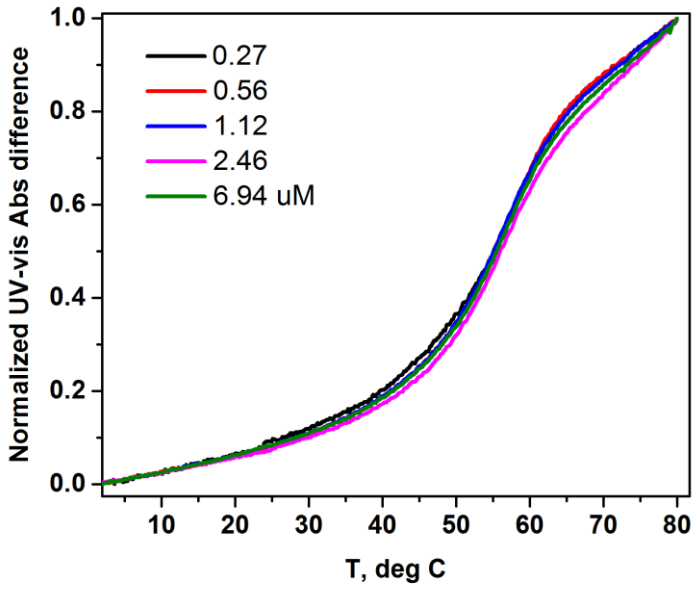

Figure 20. Overlay of normalized UV-vis melting data for $(CAGAGG)_{10}$ at varying strand concentrations, shown in the legend. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

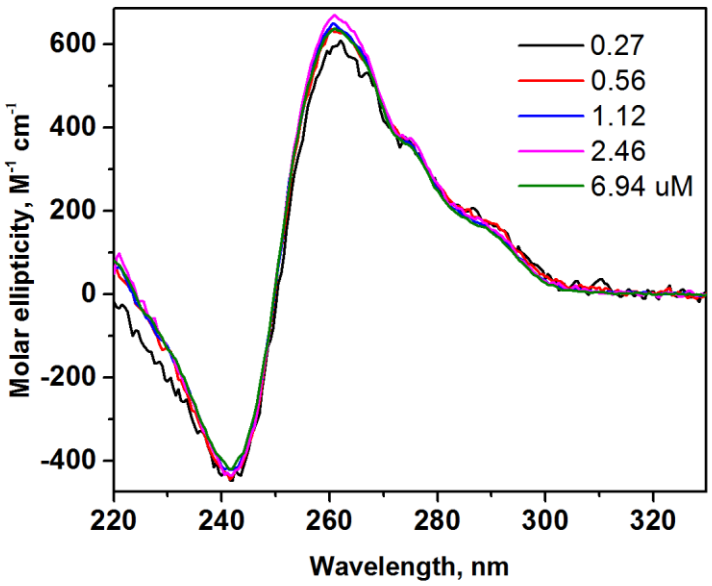

Figure 21. CD scans at 4°C for $(CAGAGG)_{10}$ at varying strand concentrations post UV-vis melting. Experiments were completed in 100K2Mg buffer (10 mM lithium cacodylate pH 7.2, 100 mM KCl and 2 mM $MgCl_2$).

Collectively, these data demonstrate that the repeats most commonly found within RPLs are the same ones that form the most stable secondary structure independently *in vitro*.

2.4 Simple tandem repeats in RPLs lead to fork stalling *in vitro* and *ex vivo*

To investigate whether the structure-forming repeat sequences enriched in RPLs were sufficient to cause replicative polymerase pausing, an *in vitro* primer extension assay was performed that models DNA synthesis on the lagging strand of the replication fork. Single-stranded DNA templates, containing either the $(CAGAGG)_{15}$ tract or its pyrimidine-rich complement, $(CCTCTG)_{15}$, were incubated with the recombinant four-subunit human DNA polymerase δ holoenzyme (Polδ4/PCNA/RFC; Pol δHE) complex, which is loaded adjacent to a radiolabeled primer interface by the RFC1-5 complex. After nascent stand

extension that was primed 68 nucleotides upstream of the inserts, products were separated on a denaturing polyacrylamide gel and visualized by autoradiography (Figure 22). Defects in nascent strand extension are indicated by the accumulation of products at the sites of polymerase obstruction.

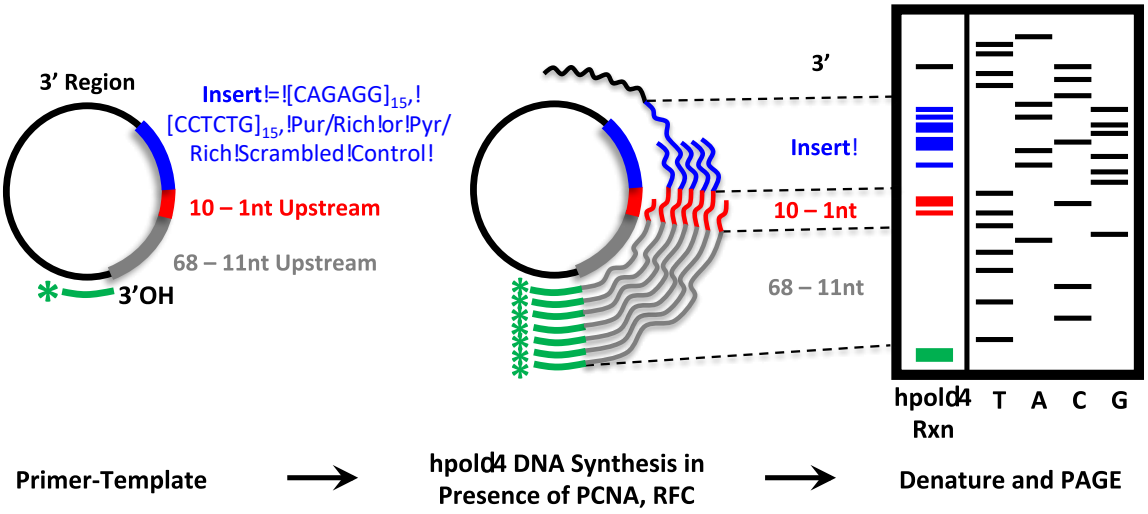

Figure 22. Schematic of *in vitro* primer-extension assay. ssDNA templates containing repeat inserts [$(CAGAGG)_{15}$ or $(CCTCTG)_{15}$] or scrambled control inserts [purine-rich or pyrimidine-rich] are hybridized to a radiolabeled primer, initiating DNA synthesis 68 nt upstream of the inserts. Pol δHE DNA synthesis products are separated by denaturing polyacrylamide gel electrophoresis, alongside a dideoxynucelotide sequencing ladder generated from the same template. After Phosphorimager analyses, DNA sequences that are inhibitory to Pol δHE synthesis are identified by the increased accumulation of reaction products at a specific position.

Notably, the purine-rich strand (CAGAGG) caused substantial accumulation of reaction products directly adjacent to the repeats insertion site, indicating pausing of the Pol δHE complex directly at the interface with the $(CAGAGG)_{15}$ repeat template (Figure 23). This outcome was contrasted by smooth progression of DNA synthesis over the pyrimidine-rich template (CCTCTG), as indicated by the absence of terminated reaction products.

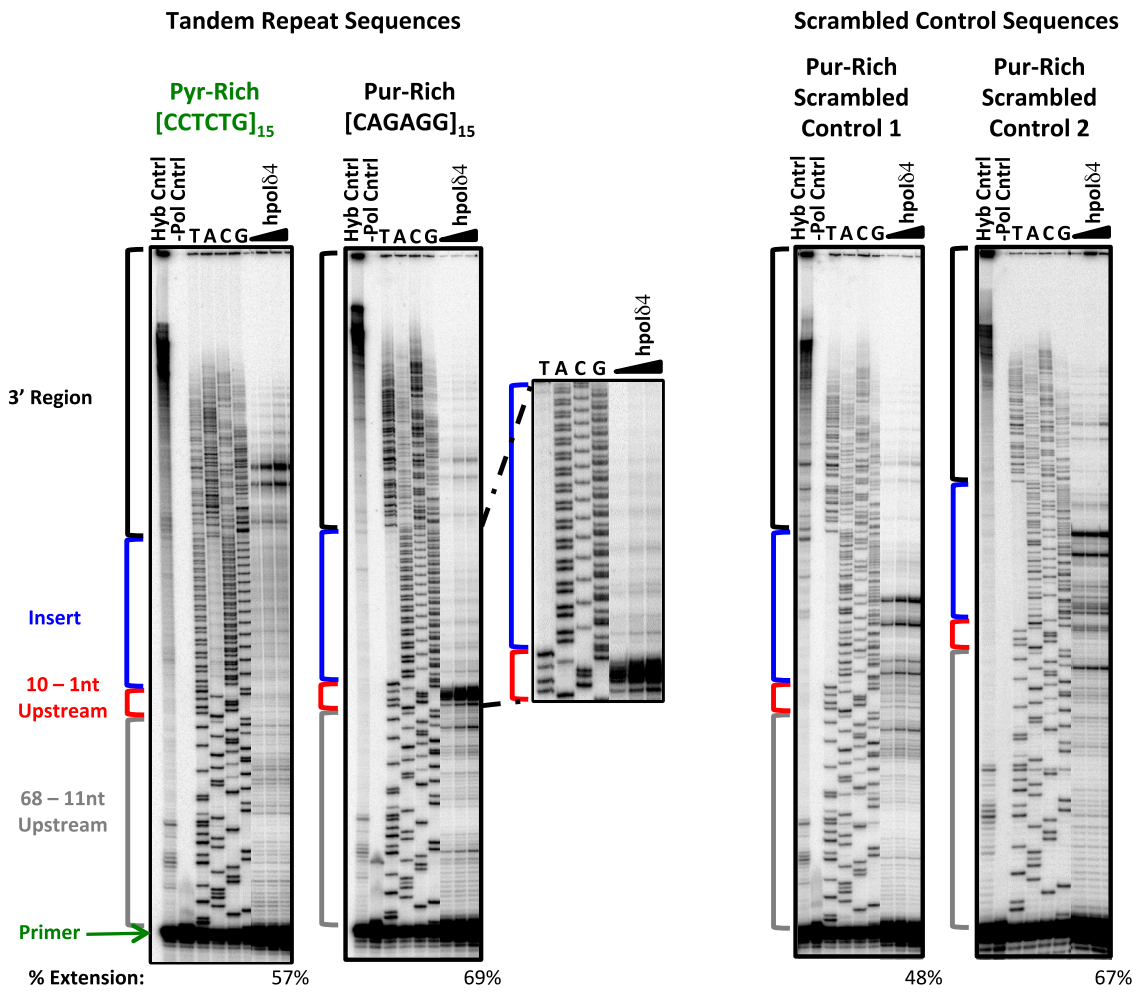

Figure 23. Representative images of Pol δHE reaction products. *Left panel:* $(CCTCTG)_{15}$ and $(CAGAGG)_{15}$ insert-containing templates; *Right panel,* two separate purine-rich scrambled control insert-containing templates. Control lanes are indicated (-Pol, No Polymerase; Hyb, Primer-template hybridization. TACG, Dideoxy sequencing ladder. Percent Extension is the total number of extended DNA molecules divided by extended molecules plus unextended primer molecules. Triangle represents increasing Pol δHE reaction time (3 – 15 minutes). Numbers 1 – 5 on the right panels indicate positions of Pol δHE pause sites within purine-rich scrambled control inserts that correspond to short hairpin structures.

Termination probabilities, normalized by sequence length, were calculated to quantify the extent of pausing at different regions among the different templates under

similar Pol δHE percent extensions. Termination of the Pol δHE complex immediately upstream of the $(CAGAGG)_{15}$ insert was 22-fold greater than that observed immediately upstream of the complementary $(CCTCTG)_{15}$ repeat, and 9- to 12-fold greater than that of the purine-rich scrambled controls (Fig. 5C; $p<0.0001$, 2 way ANOVA) (Figure 24). Notably, we observed no increase in Pol δHE termination within the repetitive $(CAGAGG)_{15}$ or $(CCTCTG)_{15}$ inserts themselves (Figure 24). To determine if these different outcomes were due solely to purine and pyrimidine base enrichments in these templates, the nucleotide sequence of repeat-containing templates were scrambled to generate two purine-rich and two pyrimidine-rich controls and tested similarly. In all cases, an increase in termination at the interface of the purine-rich insertion was not observed (Figures 23 and 24). These data indicate that CAGAGG repeats, but not the CCTCTG complementary repeat, is sufficient to stall the Pol δHE complex (Figures 23 and 24), which correlates with the relative structure-forming properties the corresponding oligos.

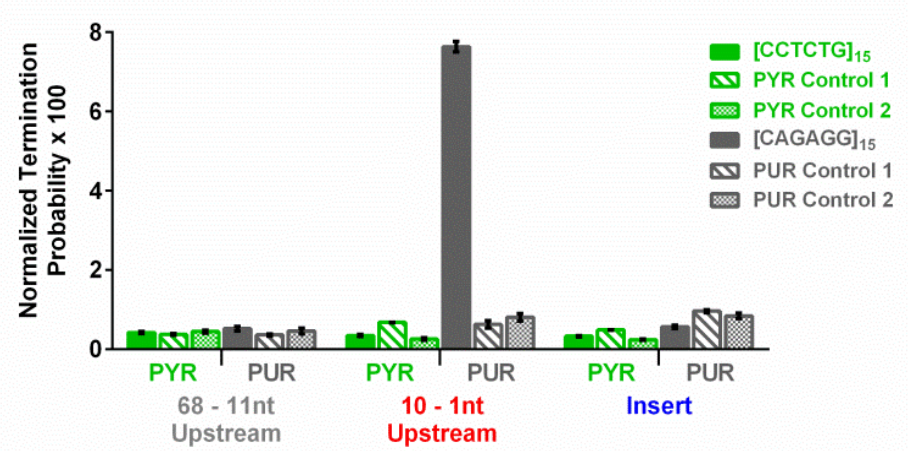

Figure 24. Pol δHE termination probability. Termination probability was quantitated after ImageQuant analyses of Phosphorimager scans, and is defined as the ratio of the number of DNA molecules within a region of interest to this number plus all longer DNA molecules. Termination probability was normalized to the number of nucleotides in each region. The data are represented as mean +/- SEM of three independent polymerase reactions for each template.

To measure the ability of the CAGAGG repeats to impede DNA replication *ex vivo*, a perfect 105 repeat stretch of CAGAGG was amplified from a RPL peak at chr7:35944716-35946239 (mm9 coordinates) and subcloned into the pML113 plasmid system (Follonier et al., 2013) at both SV40 origin-proximal and origin-distal sites and in both orientations (Figure 25). A scrambled CAGAGG synthetic sequence was generated and similarly subcloned as a control. These plasmids were then transfected into TAg-expressing U2OS cells to stimulate replication from the SV40 origin. Replicated plasmid DNA (Dpn I-resistant) was isolated and replication intermediates were resolved by neutral-neutral 2D gel electrophoresis with Southern blot detection (Figure 25).

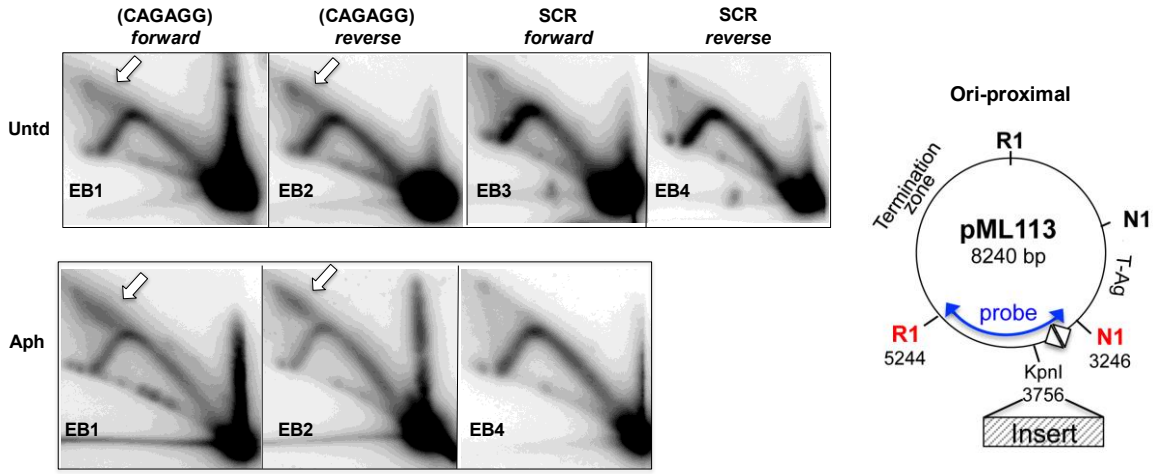

Figure 25. 2D gel of replication intermediates arising from replication through $(CAGAGG)_{105}$ in ori-proximal vector. *Right:* Schematic of the ori-proximal vectors. The pML113 parent vector encodes the bidirectional SV40 origin (triangles) and the SV40 large T-antigen (Tag), which are sufficient to support episomal plasmid replication in human cells (Follonier et al., 2013). The endogenous termination zone has been mapped between the EcoRI sites. A 630 bp cassette containing $(CAGAGG)_{105}$ tandem repeats was cloned in two orientations (forward/reverse) relative to the origin. As controls, a scrambled sequence of the same nucleotide composition and length (SCR) was similarly cloned in both orientations. *Left:* Representative 2D gels. Cells were transfected with ori-proximal vectors, and either untreated (none) or treated with 0.6μM aphidicolin (Aph) after 24 hours. Episomal DNA was isolated 48 hours after transfection. Replicated DNAs were digested

with DpnI, EcoRI (RI) and Eco NI (NI). Purified replication intermediates were separated by 2D neutral-neutral gel electrophoresis, followed by transfer to a nylon membrane and Southern hybridization using the indicated probe from plasmid pML113 (no insert). Arrows denote accumulated double-Y structure intermediates in repeat-containing vectors.

2D gel electrophoresis is a method to separate out replication intermediates based on the size and shape of the structures formed at the replication fork. Replicated DNA that has been purified and digested are first run out on an agarose gel and separated by mass, or how far replication has proceeded (1n to 2n), utilizing a low percentage agarose and low voltage. Once resolved by molecular weight, the DNA is then run out on a second dimension gel that separates the samples by structure, utilizing a high percentage agarose, high voltage, and high concentration of ethidium bromide. This increases the mobility of the samples and thus resolution by molecular shape as the DNA migrates through the gel. Replication structures related to the replication fork proceeding through a particular DNA sequence are often observed by this method as Y arcs, bubble arcs, or double-Y arcs.

2D gel results demonstrated that repeats inserted at origin-proximal sites led to the generation of migration products that were consistent with double-Y structures (Figure 25). Specifically, distinct replication intermediates were observed emanating from the top of the simple Y arc from either (CAGAGG) repeat orientation (Figure 25). However, no accumulation of products was observed on the ascending arm of the Y at the position of the repeat inserts. Because the replication termination zone (where the left and right forks converge) of the pML plasmids is present between the EcoRI sites (Follonier et al., 2013), we interpret the aberrant products to be double-Y intermediates, generated by replication fork stalling on both sides of the inserted repeat region (Huberman, 1997). In this case, slowing or stalling of the left fork before or within the repeats results in the right fork

traveling past the natural RI-RI termination zone and terminating instead within the fragment analyzed (Figure 26). Therefore, the simple Y pattern is converted to a double-Y pattern, with forks present at both ends of the fragment. Treatment of transfected cells with 0.6μM aphidicolin for 24 hours increased the abundance of double-Y migration products in the repeat-containing vectors. (Figure 25).

In contrast, simple Y structures were observed for replication through the scrambled control inserts, as expected, even after aphidicolin treatment of transfected cells to induce replication stress (Figure 25).

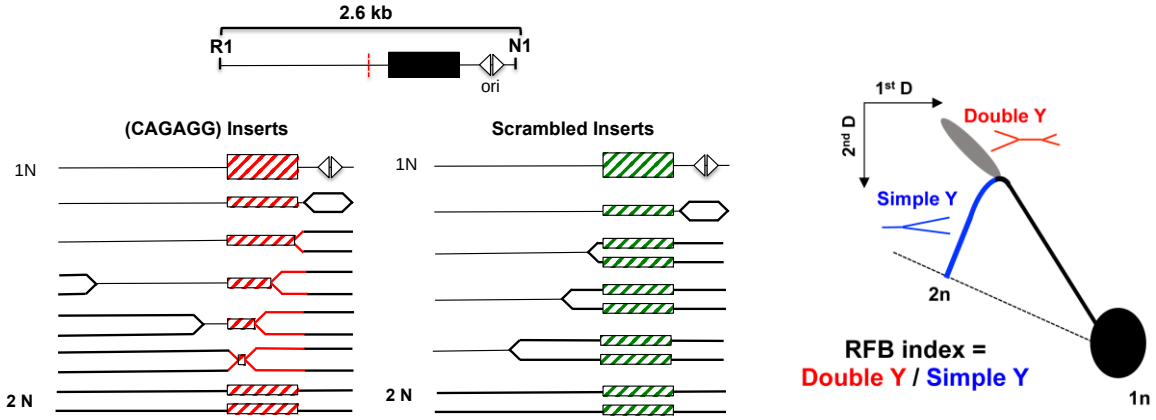

Figure 26. Schematic of replication through ori-proximal vectors. Top cartoon depicts position (203 bp) of (CAGAGG) or scrambled control inserts relative to the bidirectional origin of replication. Dashed red line indicates the center of the RI-NI fragment (1.5N size), the expected apex of a simple Y arc. Right cartoon illustrates replication fork barrier (RFB) index quantitation. The index is defined as the number of partially replicated DNA molecules present within double Y structures (red) divided by the number present in >1.5N simple Y structures (blue).

To quantitate this novel replication pattern, we calculated the replication fork barrier (RFB) index, defined as the number of partially replicated (>1.5N) DNA molecules present within double-Y structures divided by the number of partially replicated molecules present

in simple Y structures (Figure 27). The presence of the CAGAGG repeats increased the RFB index two- to three-fold over the corresponding scrambled control inserts, and this stalling differential is enhanced five-fold by aphidicolin-induced replication stress (Figure 27).

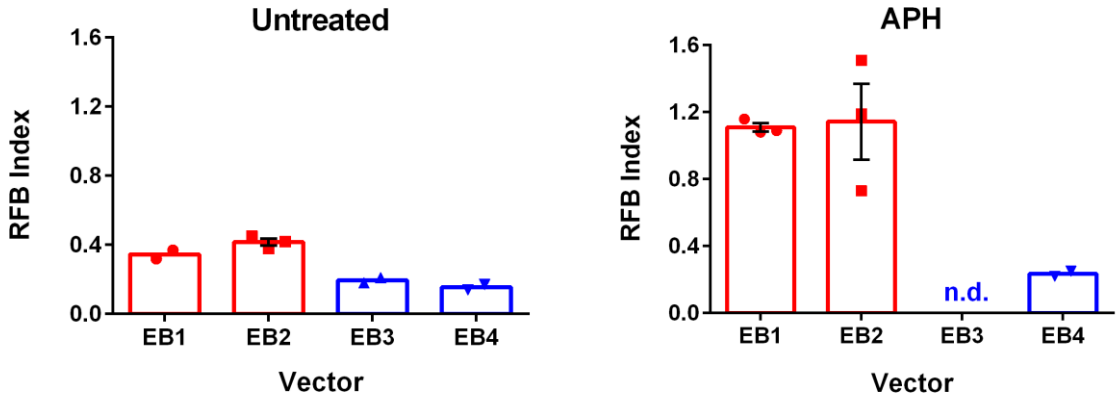

Figure 27. Quantitation of the RFB index after ori-proximal vector replication in U2OS cells. EB1 = (CAGAGG) forward; EB2 = (CAGAGG) reverse; EB3 = SCR forward; EB4 = SCR reverse. The data are represented as mean +/- SEM.

To confirm the presence of a replication fork barrier, we moved the repeats to a location 2.7 kb distant from the SV40 origin (Figure 28). Under this experimental design, the repeats would be present on the descending arm of the simple Y arc. In untreated cells, we observed a faint double-Y spike emanating from the descending arm at the position expected of the repeats. Treatment of cells with 0.6μM aphidicolin (Aph) further enhanced the presence of double-Y intermediates (Figure 28). These results are similar to those for the ori-proximal vector, and are consistent with slowing of the right fork within the repeat sequence, such that the two replication forks now terminate within the fragment analyzed.

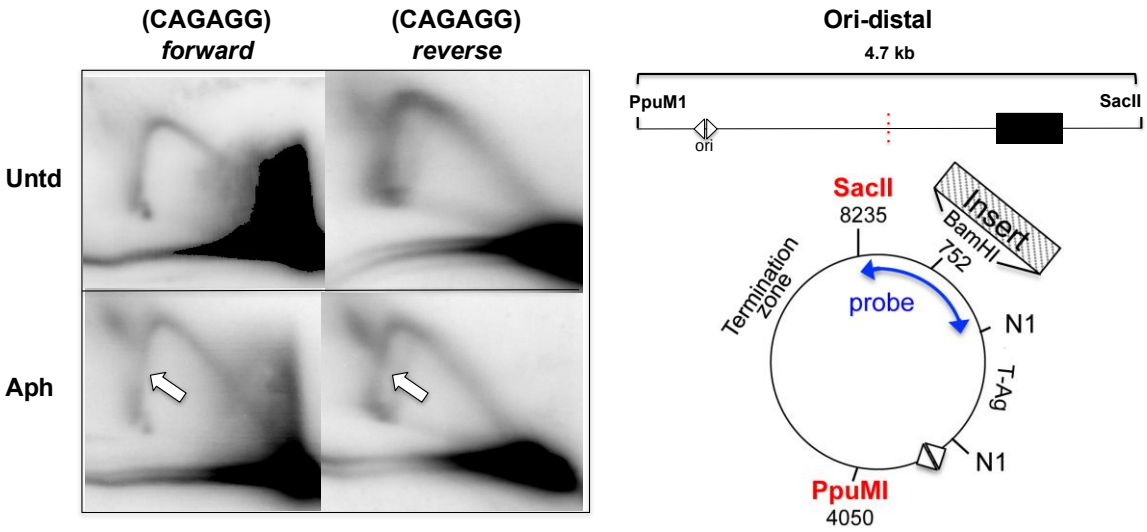

Figure 28. 2D gel of replication intermediates arising from replication through $(CAGAGG)_{105}$ in ori-distal vector. *Left:* Schematic of the ori-distal vectors. The 630 bp $(CAGAGG)_{105}$ cassette was cloned at the BamHI site, in two orientations (forward/reverse). In this vector, the repeats are located 2.7 kB from the origin. *Right:* Representative 2D gels. The experiment was carried out as described in (A), except that the purified DNAs were digested with DpnI, PpuMI, and SacII, and the SacII to EcoNI fragment from pML113 was used as a probe.

Altogether, the results of these experiments further corroborated that CAGAGG repeats are sufficient to cause replication fork stalling.

In summary, RPA ChIP-Seq results identified 173 distinct regions of significant RPA accumulation, or RPLs, that were specific to replication stress conditions caused by ATRi and low-dose aphidicolin treatment. By a program developed to accurately quantify enrichment of diverse repeats in RPA ChIP-Seq reads (REQer), previously uncharacterized simple repeat sequences were discovered to be enriched and prevalent among these sites of significant RPA accumulation. The most frequent of these repeats, CAGAGG, was shown to form a highly stable and unique intramolecular secondary structure by various structural assays, suggesting a role for structure formation in fork collapse. Tandem occurrence of this simple repeat was further proven to be sufficient to

pause the polymerase and stall progressing replication forks both *in vitro* and *in vivo*. Thus, by employing RPA ChIP-Seq on mouse cells experiencing replication stress through aphidicolin treatment in the absence of a functional ATR response pathway, we have been able to reveal a genome-wide view of regions most prone to fork stalling and collapse. These fork collapse sites were demonstrated to occur prevalently at regions of dense CAGAGG and CACAG simple repeat regions, indicating these sites to be most difficult to replicate under these conditions.

References

Aguilera A, García-Muse T. (2013). Causes of genome instability. Annu Rev Genet. *47*, 1-32.

Barlow JH, Faryabi RB, Callén E, Wong N, Malhowski A, Chen HT, Gutierrez-Cruz G, Sun HW, McKinnon P, Wright G, Casellas R, Robbiani DF, Staudt L, Fernandez-Capetillo O, Nussenzweig A. (2013). Identification of early replicating fragile sites that contribute to genome instability. Cell. *152*, 620-632.

Brewer BJ, Fangman WL. (1988). A replication fork barrier at the 3' end of yeast ribosomal RNA genes. Cell. *55*, 637-643.

Brown EJ, Baltimore D. (2000). ATR disruption leads to chromosomal fragmentation and early embryonic lethality. Genes Dev. *14*, 397-402.

Brown EJ, Baltimore D. (2003). Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. Genes Dev. *17*, 615-28.

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Cañizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, De Michele G, Filla A, De Frutos R, Palau F, Patel PI, Di Donato S, Mandel JL, Cocozza S, Koenig M, Pandolfo M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. *271*, 1423-1427.

Charrier JD, Durrant SJ, Golec JM, Kay DP, Knegtel RM, MacCormick S, Mortimore M, O'Donnell ME, Pinder JL, Reaper PM, Rutherford AP, Wang PS, Young SC, Pollard JR. (2011). Discovery of potent and selective inhibitors of ataxia telangiectasia mutated and Rad3 related (ATR) protein kinase as potential anticancer agents. J Med Chem. *54*, 2320-2330.

Cimprich KA, Cortez D. (2008). ATR: an essential regulator of genome integrity. Nat. Rev. Mol. Cell Biol. *9*, 616–627.

Flynn RL and Zou L. (2011). ATR: a master conductor of cellular responses to DNA replication stress. Trends Biochem Sci. *36*, 133-140.

Follonier C, Oehler J, Herrador R, Lopes M. (2013). Friedreich's ataxia-associated GAA repeats induce replication-fork reversal and unusual molecular junctions. Nat Struct Mol Biol. *20*, 486-494.

Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkert AJ, Holden JJ, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell. *67*, 1047-1058.

Ghamrasni SE, Cardoso R, Li L, Guturi KK, Bjerregaard VA, Liu Y, Venkatesan S, Hande MP, Henderson JT, Sanchez O, Hickson ID, Hakem A, Hakem R. (2016). Rad54 and Mus81 cooperation promotes DNA damage repair and restrains chromosome

missegregation. Oncogene. *35*, 4836-4845.

Gruber M, Wellinger RE, Sogo JM. (2000). Architecture of the replication fork stalled at the 3' end of yeast ribosomal genes. Mol. Cell. Biol. *20*, 5777-5787.

Harrigan JA, Belotserkovskaya R, Coates J, Dimitrova DS, Polo SE, Bradshaw CR, Fraser P, Jackson SP. (2011). Replication stress induces 53BP1-containing OPT domains in G1 cells. J Cell Biol. *193*, 97-108.

Hoffman EA, McCulley A, Haarer B, Arnak R, Feng W. (2015). Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription. Genome Res. *25*, 402-412.

Huberman JA. (1997). Mapping replication origins, pause sites, and termini by neutral/alkaline two-dimensional gel electrophoresis. Methods. *13*, 247-257.

Lahiri M, Gustafson TL, Majors ER, Freudenreich CH. (2004). Expanded CAG repeats activate the DNA damage checkpoint pathway. Mol Cell. *15*, 287-293.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813-1831.

Letessier A, Millot GA, Koundrioukoff S, Lachages AM, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M. (2011). Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature. *470*, 120-123.

Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LP. (2001). Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science. *293*, 864-867.

MacDougall CA, Byun TS, Van C, Yee MC, Cimprich KA. (2007). The structural determinants of checkpoint activation. Genes Dev. *21*, 898-903.

Mandel JL, Heitz D. (1992). Molecular genetics of the fragile-X syndrome: a novel type of unstable mutation. Curr Opin Genet Dev. *2*, 422-430.

Mergny JL, Li J, Lacroix L, Amrane S, Chaires JB. (2005). Thermal difference spectra: a specific signature for nucleic acid structures. Nucleic Acids Res. *33*, e138.

Minocherhomji S, Ying S, Bjerregaard VA, Bursomanno S, Aleliunaite A, Wu W, Mankouri HW, Shen H, Liu Y, Hickson ID. (2015). Replication stress activates DNA repair synthesis in mitosis. Nature. *528*, 286-290.

Sfeir A, Kosiyatrakul ST, Hockemeyer D, MacRae SL, Karlseder J, Schildkraut CL, de

Lange T. (2009). Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. Cell. *138*, 90-103.

Shechter D, Costanzo V, Gautier J. (2004). Regulation of DNA replication by ATR: signaling in response to DNA intermediates. DNA Repair. *3*, 901-908.

Smith KD, Fu MA, Brown EJ. (2009). Tim-Tipin dysfunction creates an indispensable reliance on the ATR-Chk1 pathway for continued DNA synthesis. J Cell Biol. *187*, 15-23.

Szilard RK, Jacques PE, Laramée L, Cheng B, Galicia S, Bataille AR, Yeung M, Mendez M, Bergeron M, Robert F, Durocher D. (2010). Systematic identification of fragile sites via genome-wide location analysis of gamma-H2AX. Nat Struct Mol Biol. *17*, 299-305.

Weitao T, Budd M, Hoopes LL, Campbell JL. (2003). Dna2 helicase/nuclease causes replicative fork stalling and double-strand breaks in the ribosomal DNA of Saccharomyces cerevisiae. J Biol. Chem. *278*, 22513-22522.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D, Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L, Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B. (2014). Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. Nature. *515*, 355-364.

Zegerman P, Diffley JF. (2009). DNA replication as a target of the DNA damage checkpoint. DNA Repair. *8*, 1077-1088.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zou L, Elledge SJ. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. Science. *300*, 1542–1548.

CHAPTER 3: BrITL ON REPLICATION-STRESSED MEFs

3.1 Development and validation of BrITL, DNA break-detection assay

As previously described, RPA molecules accumulate on ssDNA present at destabilized forks that maintain intact parental strands and at forks that get cleaved and processed into resected breaks. Thus, retrieving RPA-bound sites through RPA-ChIP cannot differentiate between stalled replication forks and resected DSBs. Because the 3' hydroxyl at the terminal base of blunt and resected DSBs may be relatively accessible in comparison to native chromatin, we reasoned that a combination of cell permeabilization immediately after cell harvest and biotin-end labeling by terminal deoxynucleotidyl transferase (TdT) could be sufficient to label DSB ends for retrieval. This method would demonstrate that RPL sites are prone to breakage, indicative of fork collapse sites becoming vulnerable to fork-processing events that lead to cleavage in the absence of ATR's protective function. To address this hypothesis, we developed a method termed BrITL (Break Identification by TdT Labeling), which labels and retrieves genomic regions at the site of DNA breakage.

BrITL is an assay that aims to detect de novo DNA breaks genome-wide while minimizing artefactual breaks that can arise from cell lysis and genomic extraction by labeling DNA breaks directly in permeabilized cells. In this assay, ~2-3 million cells with induced replicative stress and accumulated DNA damage are harvested, permeabilized, and incubated in a reaction mixture containing biotin-16-ddUTP in the presence of TdT, which catalyzes the addition of the biotinylated nucleotide to the ends of double-strand break fragments, with a preference for 3' overhangs. By labeling breaks within cells directly (prior to lysis and treatment with Proteinase K), we minimize the occurrence of passive and induced breaks during processing and are able to obtain a much higher

signal-to-noise ratio of endogenous break sites. After extraction of genomic DNA from the cells, DNA is sonicated to a size range that dictates the resolution of break sites (0.2-2 kb). By allowing the upper limit of fragment sizes to be 2 kb, we increase the likelihood that repetitive DNA sequences can be mapped to the reference genome through unique sequences adjacent to the repeats. Biotin-labeled DNA ends are next selected by interaction with streptavidin-coated beads in a binding buffer that greatly enhances attachment of kilobase-size DNA fragments to the streptavidin. After washes to remove non-specific binding, labeled DNA fragments are eluted from the beads by treatment with Proteinase K and SDS. Isolated DNA is further purified for analysis by qRT-PCR or prepared into a library for next-generation sequencing (Figure 29). In this manner, DSBs are tagged *in vivo* and pulled down for downstream sequence analysis.

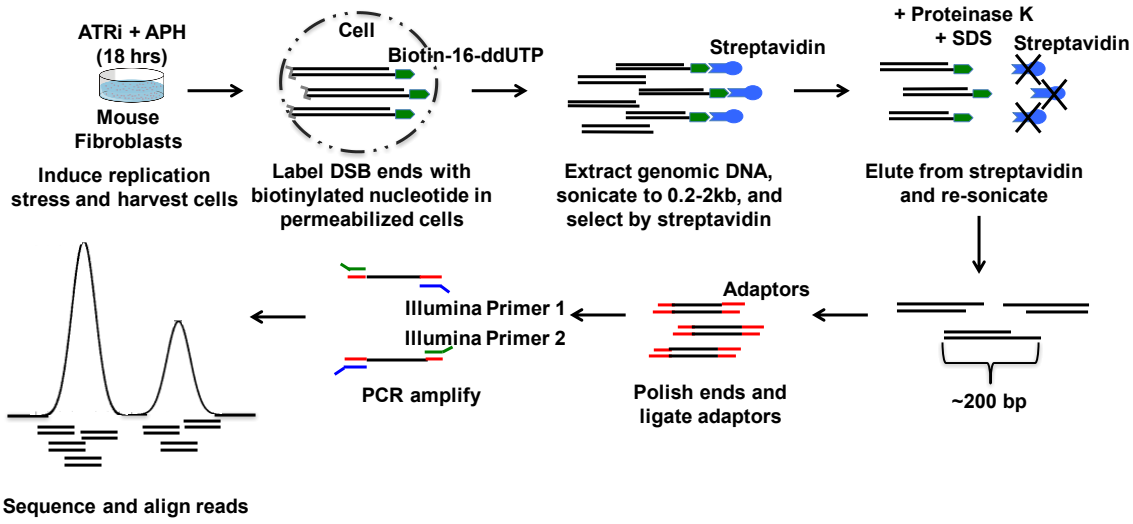

Figure 29. Schematic of BrITL method. Cells treated with 1 μM ATRi and 0.2 μM aphidicolin for 18 hours (ATRi+aph^{18hrs}) are harvested and incubated in a reaction mixture with terminal deoxynucleotidyl transferase (TdT) and biotin-16-ddUTP. Genomic DNA containing the labeled DNA ends are extracted from the cell and sonicated to a size range between 0.2-2 kb. Labeled fragments are selected by binding to streptavidin-coated beads. Non-labeled DNA fragments are washed off. Biotinylated DNA fragments are eluted from the beads by treatment with Proteinase K and SDS. DNA is purified and re-

sonicated to 200 bp. DNA ends are polished, adaptor-ligated and PCR-amplified with Illumina primers into a library for next-generation sequencing through the Illumina HiSeq platform. Reads generated from 100 bp single-end sequencing run are aligned to the reference mouse genome.

Validation by I-PpoI cleavage:

To validate BrITL as a sensitive break-detection assay, a generated site-specific break in the genome served to test BrITL's capacity to label and isolate the break region from cells. I-PpoI is a restriction enzyme with a cleavage recognition sequence in the 45S rDNA repeats, which contain about 200 copies in the genome (Gibbons et al., 2015). Murine embryonic fibroblasts were engineered to express the I-PpoI endonuclease fused to an unstable FKBP12 mutant (Goldstein et al., 2013) and a modified form of the estrogen receptor (ERT2) that binds specifically to 4-hydroxytamoxifen (4-OHT). The FKBP12-derived destabilization domain targets the protein for degradation unless bound to the FKBP12 Shield-1 ligand. The inclusion of the ERT2 domain causes the expressed protein to translocate to the nucleus upon binding to 4-OHT. This system thus allowed conditional nuclear expression of the I-PpoI restriction enzyme through combinatorial treatment of Shield-1 and 4-OHT (Figure 30). Cells from the parental line and cells expressing the construct were treated with 1 µM Shield-1 and 0.5 µM 4-OHT for 14 hours to stabilize and activate the I-PpoI fusion protein before collection for BrITL analysis.

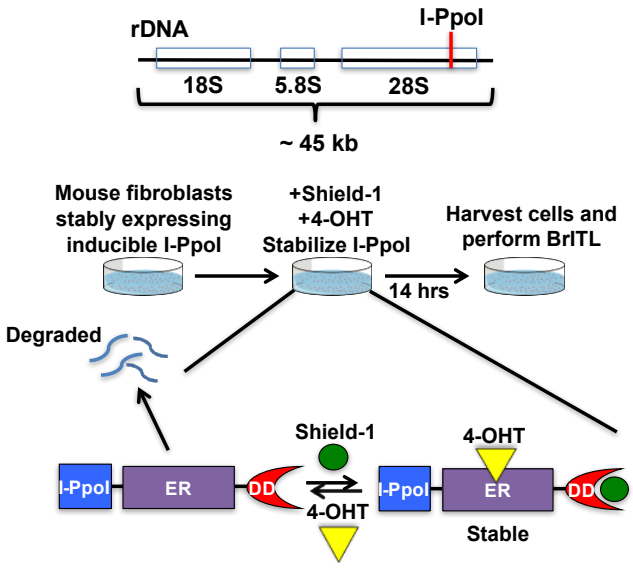

Figure 30. Experimental schematic of induced site-specific break at I-PpoI site. A break induced at the I-PpoI cleavage sequence (designated by the red line) within rDNA repeats serves as a read-out for the ability of BrITL to label and isolate DNA DSB ends. Cells were transduced with a construct conditionally expressing the I-PpoI restriction enzyme through Shield-1 stabilization and 4-hydroxytamoxifen (4-OHT)-mediated nuclear localization. Cells either expressing or not expressing the construct were treated with 1 µM Shield-1 and 0.5 µM 4-OHT for 14 hours before being harvested for BrITL.

Genomic sites that surround endogenous mammalian I-PpoI sites were then quantified in biotin-DNA retrievals by qRT-PCR as a percent of total DNA present in the input material. Primers were designed for regions at increasing distances from the I-PpoI cleavage site on the rDNA sequence. BrITL qRT-PCR results demonstrated a substantial increase in detection of genomic DNA nearest the I-PpoI cleavage site in biotin-DNA retrievals from cells in which the I-PpoI fusion protein was induced (Figure 31). The amount of I-PpoI DNA fragments retrieved by this method represented approximately 3% of the total number of I-PpoI sites present in the starting material, which includes in rDNA sequences within silenced chromatin and detected despite competent DSB repair (Figure 31). In comparison, biotin-DNA retrievals from untreated cells and those not expressing

the I-PpoI fusion protein exhibited no significant enrichment of DNA at these I-PpoI proximal regions (Figure 31). Regions 20 kb away from the I-PpoI endonuclease site were also not readily detected, even in I-PpoI-induced cells (Figure 31). These data indicate that this native chromatin break-detection method is capable of retrieving DSB ends efficiently and specifically.

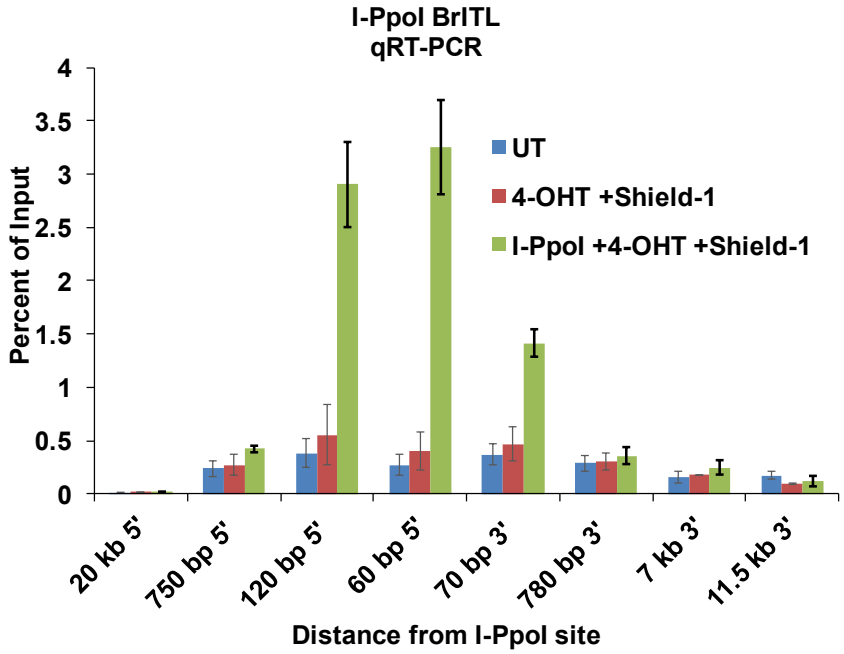

Figure 31. BrITL qRT-PCR of genomic regions surrounding the I-PpoI site. Primer sets were used at specified distances from the I-PpoI cleavage site on the rDNA fragment relative to the start of transcription, denoted on the x-axis. The y-axis quantifies the biotin-labeled fragments pulled down, represented as the percent of input. Samples include UT (DMSO treatment of parental cell line), 4-OHT + Shield-1 (4-OHT + Shield-1 treatment of parental cell line), and I-PpoI + 4-OHT + Shield-1 (4-OHT + Shield-1 treatment of cells expressing the I-PpoI construct). The data are represented as mean +/- SEM.

3.2 Simple tandem repeats in RPLs undergo double-strand breakage

While RPA ChIP-Seq identifies sites of frequent fork collapse by the accumulation of RPA molecules at ssDNA present at uncoupled replication forks or at resected ends of

DNA breaks, the method cannot confirm the actual presence of DNA breaks at a collapsed fork. We hypothesized that fork-collapse sites are also sites of breakage due to their increased instability in the absence of a functional ATR response pathway. To examine if ATRi-induced RPA enrichment at RPLs is associated with DSBs, ATRi+aph^{18hrs} cells were analyzed by BrITL-qPCR utilizing primers designed for 4 different sites of significant RPA accumulation that contain specific repeats: $(CAGAGG/CCTCTG)_n$, $(CAGG/CCTG)_n$, $(CACAG/CTGTG)_n$, and $(CAAAA/TTTTG)_n$. As quantified by qRT-PCR, BrITL retrieved substantial levels of DNA from the chosen peaks only when cells were treated with ATRi+aph^{18hrs}, not DMSO (Figure 32). Notably, the highest levels of detection were invariably at sites adjacent to the simple tandem repeats enriched in RPLs (Figure 32). This data indicates that DSBs occur at tandem repeats of $(CAGAGG/CCTCTG)_n$, $(CAGG/CCTG)_n$, and $(CAAAA/TTTTG)_n$ specifically as a result of ATRi and aphidicolin treatment (Figure 32).

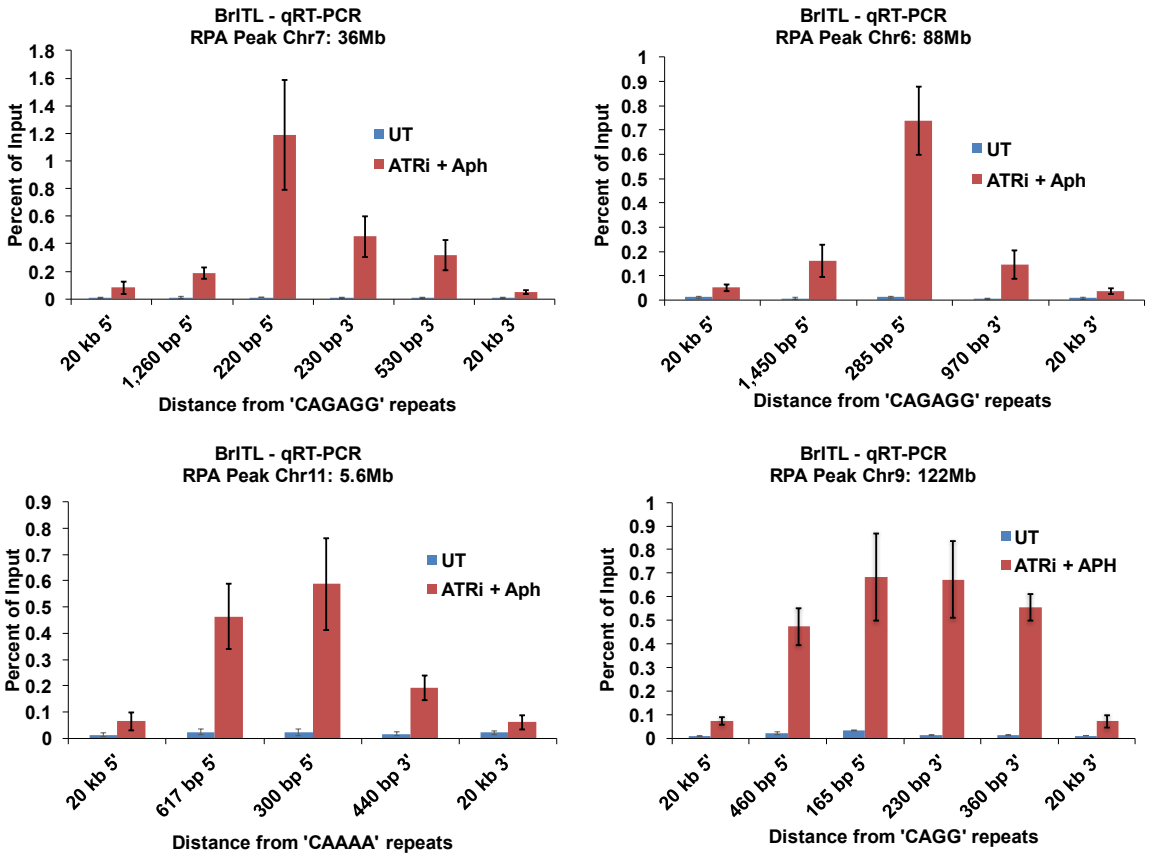

Figure 32. BrITL qRT-PCR of RPA-enriched sites centered around $(CAGAGG/CCTCTG)_n$, $(CAGG/CCTG)_n$, $(CACAG/CTGTG)_n$, and $(CAAAA/TTTTG)_n$ repeats. For each RPL site, primer sets were designed at specified distances from the central repeat region, denoted on the x-axis. The y-axis quantifies the biotin-labeled fragments retrieved, represented as percent of input. Samples include UT (DMSO treatment), and ATRi+aph^{18hrs}. The data are represented as mean +/- SEM.

One exception to these observed trends correlated with a RPL site that exhibited extensive $(CACAG/CTGTG)_n$ repeats (Figure 33). At this site, no BrITL peaks were observed adjacent to the $(CACAG/CTGTG)_n$ repeat region, although a general increase of retrieved DNA at every location tested was observed in the ATRi+aph^{18hrs} condition compared to DMSO-treated controls, indicative of an overall enhancement of instability throughout all regions assayed (Figures 32 and 33). These data indicate that some repeats may lead to RPA accumulation but do not readily cause breaks, at least ones that

are detectable by BrITL. These data suggest that RPA accumulates at specific repeat sequences that are difficult to replicate upon ATR inhibition, and break generation at these sites is frequent, but not inevitable.

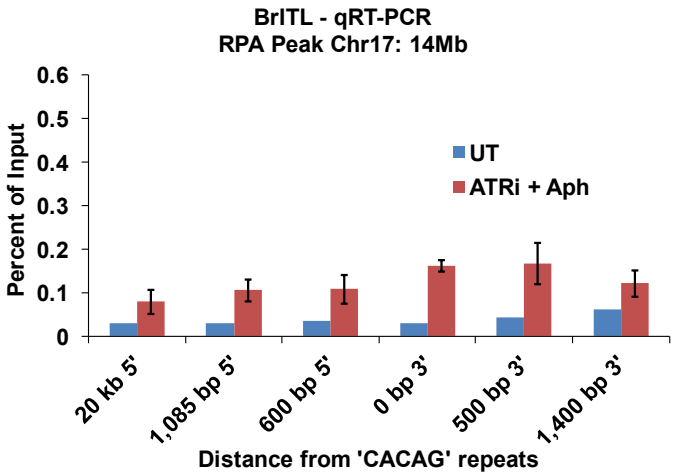

Figure 33. BrITL qRT-PCR of RPA-enriched site centered around $(CACAG/CTGTG)_n$ repeats. Primer sets were designed at specified distances from the central repeat region, denoted on the x-axis. The y-axis quantifies the biotin-labeled fragments retrieved, represented as percent of input. Samples include UT (DMSO treatment), and ATRi+aph^{18hrs}. The data are represented as mean +/- SEM.

Together, this data demonstrates that in the presence of the fork-slowing agent, aphidicolin, RPA accumulates at extensive CAGAGG/CCTCTG repeat sites that are difficult to replicate, likely due to the structure-forming properties of the purine-rich strand and the exposed ssDNA of the complimentary pyrimidine-rich strand. In the absence of the stabilizing function of ATR, these sites become increasingly susceptible to cleavage and DNA breaks. Interestingly, however, persistent breaks are not observed at CACAG/CTGTG repeat regions even though they accumulate significant amounts of RPA. The mechanisms that relate to such a stark difference between the ability of these two distinct repeats to lead to DNA breakage are still under scrutiny. One possibility is greater

freedom to form structures under the contexts of their surroundings in the genome. As indicated by the location of RPA peaks, CACAG/CTGTG repeat sites are more likely to overlap with CTCF-binding sites. It is conceivable that the presence of a protein complex surrounding the repeat sequences may protect against accessibility to structure-specific nucleases, or may even prevent the likelihood of structure formation. Structural data on these repeat sequences, discussed in the previous section, reveal that CAGAGG/CCTCTG repeats form a more stable secondary structure than CACAG/CTGTG repeats, suggesting that the amount of structure formation may play a role in break induction. Another possibility for the difference in frequency of DNA break formation stems from our previous findings on the tandem occurrences of these repeats. While CAGAGG/CCTCTG-centered RPA accumulation sites were comprised of lengthy contiguous repeat sequences, CACAG/CTGTG-centered sites were more likely to contain interruptions of intervening sequences. This may hinder perfect structure formation or decrease its stability, making it formidable for fork progression, but less likely to attract structure-specific nucleases for DNA cleavage.

3.3 BrITL-specific sites are composed of hairpin-forming inverted retroelements

BrITL applied to the ATRi+aph^{18hrs} and DMSO-treated cells next underwent deep-sequencing to reveal genome-wide sites of DNA DSBs. Following alignment to the reference genome and normalization by input DNA of BrITL reads, loci characterized by statistically significant read enrichments (>4-fold over input, p-value <10^{-3}) in both of 2 biological replicates were identified. For enrichment analysis, the biological replicates and inputs of each experimental condition underwent an IDR analysis (Landt et al., 2012) with the MACS2 peak-calling program (Zhang et al., 2008) to give the final peak list per condition. IDR thresholds of >0.05 were used for self-consistency and comparison of

biological replicates, and >0.005 for pooled-consistency analysis. Peaks that passed IDR thresholds were further filtered to select those with p-value $<10^{-3}$ and that were above 4-fold enriched over input. Regions within 2 kb of one another were merged. The final peak list per condition was generated as a set intersection with and subtraction from the DMSO-control peak list.

In aggregate, 224 peaks, or break sites, were identified by BrITL. Of those, 17 overlapped with RPL sites (~8%), 13 of which spanned extensive CAGAGG/CCTCTG repeats (Figure 34). However, the remaining BrITL peaks did not coincide with RPA-enriched regions. This suggests a mechanism of fork collapse that differs from that at break sites with abundant RPA formation. Limited RPA accumulation implies reduced amounts of ssDNA. This suggests that as a replication fork slows down through these sequences, specific structures may form on either separated duplex strand that contain minimal ssDNA, or that preclude exposure of ssDNA, yet that leads to heightened fork stalling and cleavage.

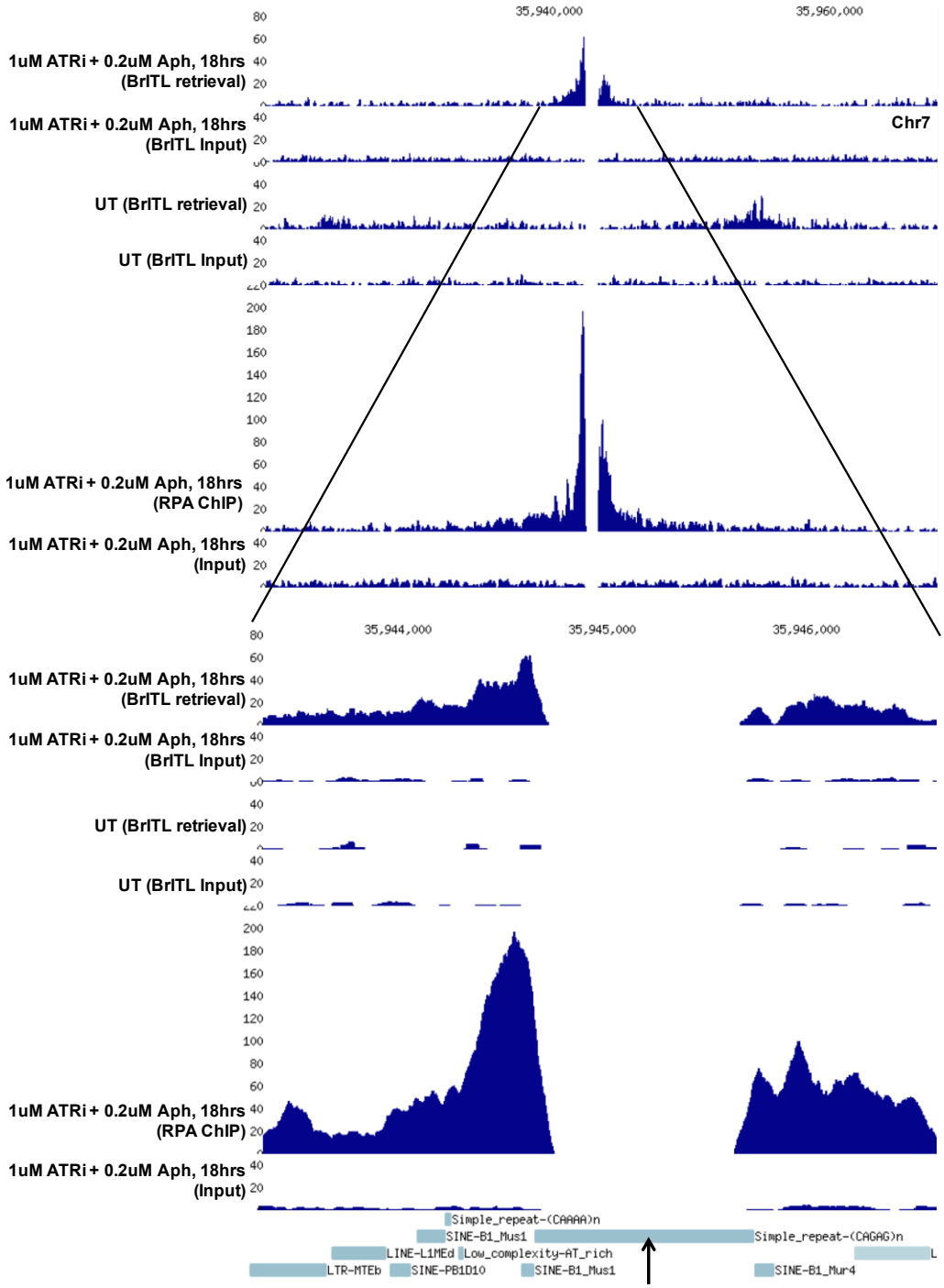

Figure 34. Overlap of BrITL and RPA peaks in the ATRi+aph^{18hrs} condition. Accumulation of reads observed in both BrITL and RPA ChIP-Seq coverage tracks of ATRi+aph^{18hrs} retrievals but not in the respective inputs or untreated controls indicates treatment-specific RPA enrichment and breakage. This site contains a region marked by an absence of reads that spans tandem CAGAG repeats (denoted by black arrow).

Consistent with this, of the 224 break sites, 147 overlapped with inverted repeats that have the potential to form hairpins, constituting ~66% of the total break sites and 71% of RPA-low break sites. M-fold, an online program that predicts DNA secondary structure and its melting temperature based on input sequences, displayed these inverted elements to form hairpin structures with an average melting temperature of 79°C at physiological salt concentrations of 135 mM Na^+ and 1 mM Mg^{2+} (Zuker, 2003).

Inverted repeats that on a duplex DNA form cruciform structures have been shown to influence various biological processes, including DNA replication and transcription. These structures are sites of binding by histones and HMG proteins, critical components of chromatin architecture (Rampakakis et al., 2010; Brázda et al., 2011), as well as by regulatory enzymes such as PARP-1, which can bind to promoter-specific cruciforms and affect transcriptional regulation (Potaman et al., 2005). Moreover, inverted repeats tend to occur not only near promoter regions, but also replication origins, at which altered supercoiled states from cruciform formation may aid in replication initiation (Pearson et al., 1996). Besides influencing the local supercoiled state of DNA, cruciform extrusion is energetically favored in events that generate negative supercoiling. The transition of duplex DNA to cruciform structures may also occur as certain proteins, such as PARP-1, binds to inverted repeat sequences (Chasovskikh et al., 2005). Thus, there are physiological functions for inverted sequences in the genome and their resulting non-B DNA structure.

However, long inverted sequences can induce chromosomal instability and genomic rearrangements (Wang and Leung, 2006), as evidenced by the frequent localization of inverted sequences at breakpoint junctions. Hairpins are capable of stalling replication forks. An *in vitro* study utilizing an *oriC* plasmid DNA replication system

containing a single origin and a 246 bp inverted sequence demonstrated the formation of hairpins from the sequence on both the leading and lagging strand ahead of the replicating fork (Lai et al., 2016). These structures, sensitive to the hairpin-specific endonuclease/exonucleause, SbcCD (bacterial Mre11/Rad50 homologue), were sufficient to block DNA synthesis and lead to fork pausing (Lai et al., 2016). Studies on human cell lines revealed numerous deletions in the center of palindromic AT-rich regions (PATRRs) and their increased rate of involvement in genomic rearrangements, signifying that PATRRs are fragile (Kurahashi et al., 2006). It is speculated that the weaker inter-strand bonds in the AT-rich sequence is conducive to strand dissociation, while the non-AT-rich regions at either end provides an anchor-like stability for the extruding cruciform structure, creating an intermediate that precedes DNA breakage and translocation (Kurahashi et al., 2006; Inagaki et al., 2009). It has been observed that the proportion of PATRR that forms a cruciform is linked to the amount of rearrangements that occur (Inagaki et al., 2009). Additionally, PATRRs often display size polymorphisms in humans, which affect the likelihood of secondary structure formation and translocation frequency, further demonstrating a correlation between cruciform extrusion and sensitivity to breakage and error-prone repair (Kurahashi et al., 2006).

To determine the prevalence of inverted repeats in the genome that exhibit features specific to BrITL break sites centered around such sequences (length of stem-loop is between 300-600 bp, loop region is <100 bp, and the structure folds in 135 mM Na^{+} and 1mM Mg^{2+}), all putative inverted repeats in the mouse genome matching those criteria were documented (Table 6). A total of 100,492 sites were recognized (Table 6). This data set reveals many more sites than those identified by BrITL, suggesting that the

presence of an inverted sequence is not enough to induce breakage under replication stress.

Chromosome	# of IR regions	Chromosome size (Mb)
1	6,345	195
2	7,753	182
3	4,371	160
4	7,204	157
5	7,098	152
6	5,267	150
7	6,730	145
8	5,491	129
9	5,504	125
10	4,679	131
11	7,432	122
12	4,329	120
13	3,633	120
14	3,791	125
15	4,148	104
16	3,179	91
17	4,343	95
18	2,935	91
19	2,641	61
X	3,619	142

Table 6. Inverted repeats in mouse genome. These inverted repeats are of length 300-600 bp, with loop region <100 bp, and for which the structure folds in 135 mM Na^+ and 1 mM Mg^{2+}.

However, while the incidence of inverted repeats within BrITL-specific sites was a critical finding, another key feature was the composition of the hairpin stem regions, which consisted exclusively of a pair of retroelements (SINEs, LINEs, or LTRs) from the same family, one on each stem (Figure 35). Interestingly, a major tendency towards SINE B1 or B2 elements was observed at these stem-loop sites (Figure 35).

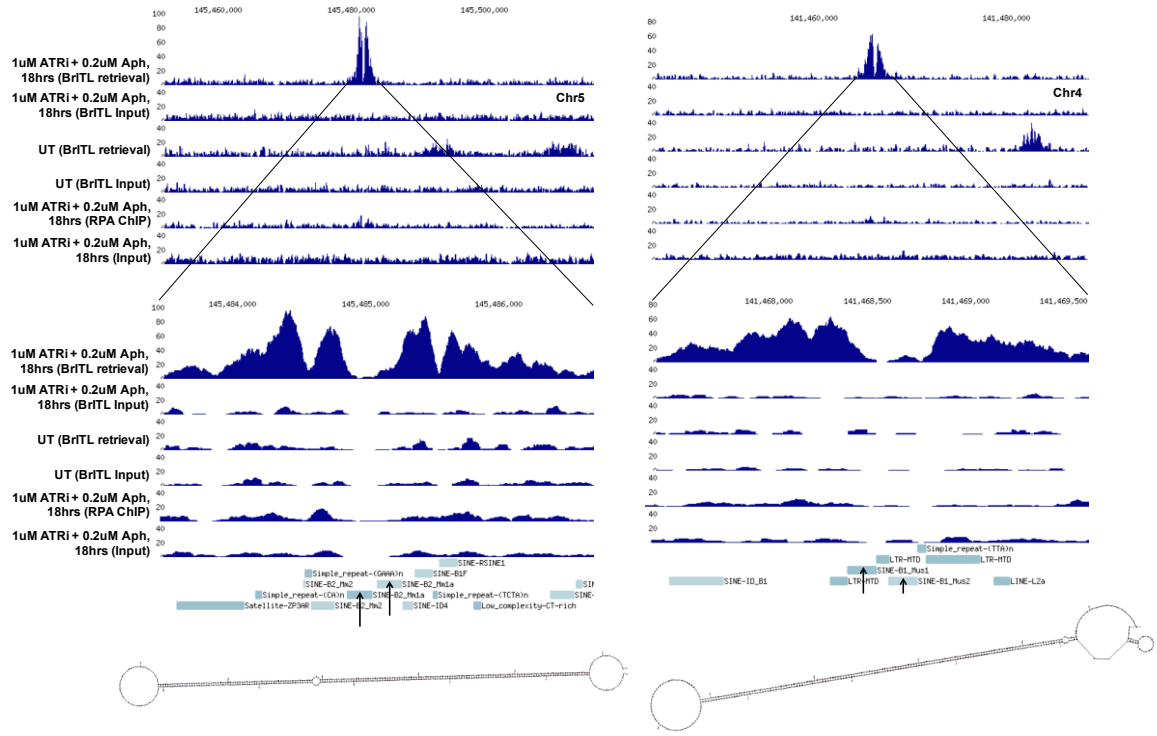

Figure 35. BrITL peaks at inverted SINEs. Accumulation of reads observed at two separate genomic regions in BrITL coverage tracks of ATRi+aph^{18hrs} but not in UT indicates treatment-specific break sites. These high-density reads are centered around regions containing a pair of SINE elements (denoted by black arrows) that form a stable hairpin structure predicted by M-fold.

In humans, only a single SINE element, the *Alu* family, is active, while mouse genomes consist of four unique SINE families: B1, B2, ID, and B4. B1 and B2 SINE elements are the major SINE families in mouse, occupying ~1.3% of the genome, with around 150,000 and 90,000 copies, respectively (Hasties, 1989). The B1 element is derived from a portion of the 7SL RNA gene and the B2 element evolved from a tRNA^{lys} gene (Hasties, 1989; Okada, 1991). The length of SINEs is relatively short: B1 is on average 140 bp long and B2 is on average 190 bp. Their short length necessitates a DNA copy of itself utilizing a reverse transcriptase transcribed from LINEs and LTRs. LINEs are similar to SINEs, but are typically 7,000 bp in length and can encode their own reverse

transcriptase. LTRs are derived from and exist on either end of a provirus, functioning to integrate the viral sequence into the host genome.

SINEs, LINEs and LTRs are mobile elements of the genome that integrate preferentially within accessible chromatin regions, capable of causing mutations in coding regions. *Alu* elements in humans have been recurrently observed at breakpoint junctions of small deletion variants (de Smith et al., 2008). But while they are prevalent in the genome, retroelements and their function remain poorly understood. Normally repressed by methylation, retroelements become activated upon cellular stress and in human diseases (Gualtieri et al., 2013). Mechanisms for their upregulation under these conditions remain unknown. However, one study discovered that increased expression of a SINE inverted repeat with sequence homology to an intron within the *BRCA1* gene downregulates *BRCA1* mRNA levels via an siRNA pathway, indicating a novel role for de-regulated SINE transcripts in silencing the expression of other genes (Peterson et al., 2013). While there is extensive knowledge about SINEs, LINEs, and LTRs in the genome, and a developing understanding of their potential roles, no study has yet focused on inverted retroelements and their presence or function in the genome.

The finding that most RPA-low break sites consists singularly of inverted retroelement sequences adds a distinguishing feature to this subset of breaks. It allows us to further filter the identified inverted repeats in the genome by excluding sites that are not generated by retroelements and for which the T_m of the resulting hairpin is <70°C. By doing so, we can determine whether inverted retroelements that generate stable hairpin structures with high melting temperatures are sufficient to induce DSBs wherever they may be in the genome under conditions of ATR inhibition and low-dose aphidicolin treatment.

However, if we find that even after such filtering, there is no correlation of the identified genomic regions to breakage, this would indicate that other factors might promote DSBs at these BrITL sites. One possibility is the role retroelement transcription may play. It has been previously observed that with cellular stress, retroelements become increasingly expressed and mobile, lending towards instability (Gualtieri et al., 2013). In addition, transcription would involve DNA duplex unwinding and exposure of ssDNA, which would provide an opportunity for these hairpins to form and to subsequently stall oncoming replication forks. It is an interesting possibility for which future studies to test such mechanisms can be explored.

Altogether, this study has provided the first demonstration of a unique class of repeats, inverted retroelements, in comprising a major fraction of DNA breaks arising from replication stress, revealing a strong inclination for these structural-prone elements to stall replication forks and cause DSBs. Of the remaining breaks that do not accumulate RPA nor form hairpins by inverted retroelements, little is currently known on the mechanisms behind their breakage. However, their characterization remains a future study.

Overall, analysis of replication-stressed MEFs by a specific DNA DSB-detection assay, BrITL, revealed that defined sites of RPA accumulation in the genome under conditions of replication stress caused by low-dose aphidicolin treatment and inhibited ATR were also sites of persistent breakage, while other sites under the same conditions were not. This suggests that certain genomic features differentially affect the level of fork breakage. Sequence composition appeared to be the most distinctive driver of fork collapse and subsequent DSB formation, as RPA-enriched sites centrally localized around CAGAGG/CCTCTG tandem repeats were more likely to result in DNA breaks. In contrast, while CACAG/CTGTG repeat regions recruited similar levels of RPA molecules under

replication stress, it was not detected by BrITL as sites of frequent breakage. This pattern suggests that the greater ability of CAGAGG/CCTCTG tandem repeats to form stable secondary structures correlates with DNA breakage, indicating that secondary structure formation plays a key role in fork collapse and DSB formation.

Surprisingly, genome-wide BrITL analysis further demonstrated the existence of a larger fraction of DNA breaks that were not accompanied by RPA accumulation, suggesting fork collapse without significant ssDNA formation. These sites were enriched for inverted retroelement sequences that are predicted to generate long, thermally stable stem-loop structures on both strands. By M-fold prediction, the average melting temperature of these structures was 79°C, indicative of great sequence homology between the inverted repeats and, thus, greater stability of the formed structure. Increased ssDNA exposure upon fork slowing with aphidicolin treatment would augment the likelihood of the extrusion of these structures, resulting in fork stalling, consistent with our ability to detect these sites under aphidicolin treatment and ATR inhibition. Why these sites tend to break more frequently than other regions is still under examination. The highly stable formation of these stem-loops, as indicated by the generally high melting temperatures, could be one driving factor. Any instance of ssDNA exposure would make it extremely likely that the inverted sequences would anneal to each other, perhaps more so than to its complimentary strand. Such instances appear during transcription or replication. These hairpin structures, in turn, would be favored substrates for various structure-specific nucleases, such as MUS81-EME1/2 and SLX4-SLX1 (Muñoz et al., 2009; Sarbajna et al., 2014; Pepe and West, 2014), resulting in heightened levels of breakage and subsequent detection by BrITL.

These findings demonstrate the discovery of a complex set of replication-sensitive regions identified genome-wide in an unbiased manner that, while it provides a plethora of novel information, also leaves us with outstanding questions. Why do only some RPA-enriched sites break, and others do not? What is the mechanism of breakage at BrITL sites that are neither enriched for RPA nor for inverted retroelements? While these sites in question consist of a minor fraction of DNA breaks under these conditions, their cause for breakage would open a new avenue of research.

References

Brázda V, Laister RC, Jagelská EB, Arrowsmith C. (2011). Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol Biol. *12*, 33.

Chasovskikh S, Dimtchev A, Smulson M, Dritschilo A. (2005). DNA transitions induced by binding of PARP-1 to cruciform structures in supercoiled plasmids. Cytometry A. *68*, 21-27.

de Smith AJ, Walters RG, Coin LJ, Steinfeld I, Yakhini Z, Sladek R, Froguel P, Blakemore AI. (2008). Small deletion variants have stable breakpoints commonly associated with alu elements. PLoS One. *3*, e3104.

Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. Proc Natl Acad Sci USA. *112*, 2485-2490.

Hasties ND. (1989). Highly repeated DNA families in the genome of Mus musculus. Genetic Variants and Strains of the Laboratory Mouse. 559-573.

Inagaki H, Ohye T, Kogo H, Kato T, Bolor H, Taniguchi M, Shaikh TH, Emanuel BS, Kurahashi H. (2009). Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. Genome Res. *19*, 191-198.

Kurahashi H, Inagaki H, Ohye T, Kogo H, Kato T, Emanuel BS. (2006). Palindrome-mediated chromosomal translocations in humans. DNA Repair. *5*, 1136-1145.

Lai PJ, Lim CT, Le HP, Katayama T, Leach DR, Furukohri A, Maki H. (2016). Long inverted repeat transiently stalls DNA replication by forming hairpin structures on both leading and lagging strands. Gens Cells. *21*, 136-145.

Muñoz IM, Hain K, Déclais AC, Gardiner M, Toh GW, Sanchez-Pulido L, Heuckmann JM, Toth R, Macartney T, Eppink B, Kanaar R, Ponting CP, Liley DM, Rouse J. (2009). Coordination of structure-specific nucleases by human SLX4/BTBD12 is required for DNA repair. Mol Cell. *35*, 116-127.

Okada N. (1991). SINEs. Curr Opin Genet Dev. *1*, 498-504.

Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M. (1996). Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. J Cell Biochem. *63*, 1-22.

Pepe A, West S. (2014). MUS81-EME2 promotes replication fork restart. Cell Reports. *7*, 1048-1055.

Peterson M, Chandler VL, Bosco G. (2013). High SINE RNA expression correlates with post-transcriptional downregulation of BRCA1. Genes. *4*, 226-243.

Potaman VN, Shiyakhtenko LS, Oussatcheva EA, Lyubchenko YL, Soldatenkov VA. (2005). Specific binding of poly(ADP-ribose) polymerase-1 to cruciform hairpins. J Mol Biol. *348*, 609-615.

Rampakakis E, Gkogkas C, Di Paola D, Zannis-Hadjopoulos M. (2010). Replication initiation and DNA topology: the twisted life of the origin. J Cell Biochem. *110*, 35-43.

Sarbajna S, Davies D, West S. (2014). Roles of SLX1-SLX4, MUS81-EME1, and GEN1 in avoiding genome instability and mitotic catastrophe. Genes Dev. *28*, 1124-1136.

Wang Y, Leung FC. (2006). Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. FEBS Lett. *580*, 1277-1284.

Zuker M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406-3415.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

Herein I have described the identification of Replication Perturbed Locations (RPLs), sensitized by exposure to low-dose aphidicolin and ATR inhibition, from two complementary genome-wide approaches. RPA ChIP-Seq recognizes sites in the genome at which forks frequently stall, retaining an intact fork structure, and at which forks get processed into resected breaks. However, as RPA molecules are recruited to ssDNA present under both conditions, RPA ChIP-Seq cannot distinguish between these two outcomes. Our second assay, BrITL, specifically retrieves sites at which stalled forks undergo frequent double-strand break formation. This method further classifies collapsed forks as sites of persistent breakage and thus regions of greater instability.

We have found that ATR inhibition does not cause replication fork collapse uniformly across the genome, but rather at highly specific sites. Importantly, sites observed using ATRi in combination with slowed polymerase progression (low dose aphidicolin treatment) were not substantially different from ATR inhibition combined with TIMELESS suppression or ATR inhibition alone. This overlap between the two different conditions indicates that helicase-polymerase uncoupling, by whatever mechanism, leads to a reliance on ATR at highly specified sites, even if that uncoupling occurs from DNA sequences that are difficult to replicate in the absence of externally enforced replication abnormalities. Fork collapse generally correlated with the ability of RPL-associated simple tandem repeats to form secondary structures *in vitro*. Notably, the most troublesome sequences found to cause replication fork collapse were not those that were described previously to form structures and impede DNA replication when expanded, but rather were repeated sequences that were less studied, if at all, for their replication effects. Finally, we

found that RPA accumulation at some of the most prevalent repeat-containing sites were prone to DSB generation, as detected by BrITL. Overall, these studies have identified a new class of difficult-to-replicate sequences that are highly dependent on ATR function for stability during DNA replication.

4.1 RPA ChIP-Seq identification of fork-collapse sites

While it is known that ATR loss leads to increased DNA damage, our results indicate that in the absence of a functional ATR pathway, RPA accumulates at specific sites throughout the genome that represent defined areas of fork collapse. In total, 173 sites of significant RPA accumulation were identified in the ATRi+aph^{18hrs} condition. The majority of these sites, termed RPLs, were characterized by extensive simple tandem repeats (92%), many of which were reflected by central regions of unaligned reads. Indeed, RPA peaks exhibiting the highest signal intensity (top 50%) were characterized by the inclusion of such simple tandem repeats (82 out of 85 peaks).

One striking feature of simple tandem repeats associated with RPLs is the lack of literature noting their instability and the relative absence of other simple tandem repeats that have been previously characterized as difficult to replicate. The one exception to this general characteristic is the CAGG/CCTG repeat found within 3 RPLs in the ATRi+aph^{18hrs} condition. The presence and expansion of this repeat within intron 1 of the human myotonic dystrophy type 2 gene (CNBP) is causative of the disease, however, the RPL-associated repeat was not observed within the mouse orthologue of this gene. According to the murine reference genome, CAGG/CCTG monomer repeats range from 11 to 618 units at associated RPLs, indicating that expansive stretches of this human microsatellite repeat are unstable following ATR inhibition. Similarly, it seems likely that triplet repeats and other sequences that have previously been shown to impede DNA replication and

cause fork collapse (Lahiri et al., 2004; Campuzano et al., 1996; Fu et al., 1991; Mandel and Heitz, 1992; Takai et al., 2003; McNees et al., 2010) upon ATR suppression might collapse if expanded beyond the threshold length associated with disease onset (Orr and Zoghbi, 2007). Our study demonstrates that a different subset of repeats in the wild-type mammalian genome are one of the causes of genomic instability following ATR inhibition.

Notably, our data indicate that fork collapse following ATR inhibition generally correlates with the ability of RPL-associated simple tandem repeats to form secondary structures. The predominant structure-forming tandem repeat identified in RPLs was $(CAGAGG)_n$, a sequence that is predicted to form only weakly stable B-type DNA structures by common programs (Zuker, 2003). However, while the TDS and CD signatures of $(CAGAGG)_n$ repeats showed strong secondary structure formation, these were not consistent with any known DNA structure according to principal component analysis of the CD signatures of over 60 previously characterized B-form or non-B-form DNA structures (personal communication, Dr. Brad Cairns, University of Louisville). Interestingly, many of the RPL-associated repeats exhibited a similar sequence pattern: C, followed by a variable intervening region, such as (AG), and ending with G (e.g. CAGG, CAGAGG, CACAG). Further biophysical characterization of these sequence-related repeats will be of interest to determine if they form similar structures.

An interesting discovery was that while $(CAGAGG)_n$ forms secondary structure and impedes DNA replication, its complementary strand $(CCTCTG)_n$ does not. This distinction implies that structure formation may only occur after repeat unwinding and will form on just one of the two strands. Because POL epsilon is associated with the MCM2-7 helicase, relatively short stretches of ssDNA will form between the N-terminal face of MCM2-7 barrel and POL epsilon. In contrast, significant lengths of ssDNA are expected to form on the

lagging strand template upon unwinding at the C-terminal front of MCM2-7 and passage over the helicase for subsequent priming and gap filling. Notably, we observed that RPL peaks exhibited asymmetric signal intensity on either side of repeat regions, implying that forks primarily encountered these repeats from the high-signal intensity side. Accordingly, the vast majority of such asymmetric peaks had $(CAGAGG)_n$ enriched on the lagging strand template (88%). While further research will be required to prove this model, these data are consistent with the dependence of fork stability on ATR due to difficulties in lagging strand synthesis.

4.2 BrITL identification of fork-collapse and break sites

BrITL analysis indicated that at some frequency, a fraction of RPLs culminates into DSBs, particularly at $(CAGAGG)_n$ repeat-containing sites. DNA breaks can occur as a potential consequence of replication fork collapse in mammalian cells (Cimprich et al., 2008). Recently, we and others demonstrated that a significant fraction of increased H2AX phosphorylation in response to ATR deletion is dependent on the SLX4-endonuclease (Ragland et al., 2013; Dungrawala et al., 2015). SLX4 forms a heterodimeric complex with SLX1 and serves as a scaffold for two other endonuclease heterodimers, MUS81-EME1 and ERCC1-XPF. This complex is fully assembled through phosphorylation of members by the CDK1-AURKA-PLK1 axis during transition into mitosis both in yeast and mammals (Sarbajna et al., 2014; Pepe and West, 2014; Szakal and Branzei, 2013; Ragland et al, 2013). Thus, ATRi-mediated replication fork collapse into breaks could simply involve the persistence of daughter strand gaps formed at replication forks followed by cleavage by prematurely activated SLX4-endonuclease complex.

Although the model above is attractive, one notable exception is the limited ability of RPLs harboring $(CACAG)_n$ repeats to be detected by BrITL. This lack of detection

occurs despite having comparable levels of RPA enrichment as at $(CAGAGG)_n$-containing RPLs, which were readily detected by BrITL. It is not clear whether the BrITL-refractory nature reflects a decreased persistence of breaks, for example, from rapid HR-mediated restart, or relates to decreased vulnerability to break formation. Regarding the latter possibility, it is conceivable that $(CAGAGG)_n$ sequences are processed into structures that are more amenable to endonuclease cleavage, such as through the formation of Holliday junctions. However, neither single nor double X-spikes were observed by 2D gel electrophoresis from $(CAGAGG)_n$-mediated fork stalling. Nevertheless, these findings indicate that RPLs can be classified into at least two distinct categories: those that break, and those that do not. The diversity of outcomes rendered by these repeats expands the complexity of fork collapse mechanisms beyond protein-mediated responses and now includes sequence content as a determining factor.

Surprisingly, it was discovered that the greater majority of break sites identified by BrITL (66%) did not accumulate significant amounts of RPA, suggesting limited availability of ssDNA at these sites of fork collapse. Most of these regions (71%) were revealed to be centrally localized around pairs of inverted retroelements (SINEs, LINEs, and LTRs) that are predicted to form highly stable hairpin structures that may be favorable substrates for cleavage by Holliday junction resolvases, such as SLX4-1 and MUS81-EME1. The classification of these sites as sources of frequent double-strand breakage from replication stress could be used to correlate regions of the genome that would be involved in genomic rearrangements and deletions under these contexts, providing evidence on the most relevant mechanisms of genomic instability that are active under these conditions. While inverted retroelements were observed in 71% of RPA-low break sites, a common characteristic or association among the remaining break regions remains undetermined.

Overall, this study has been the first to identify such elements as ATRi-sensitive and susceptible to DSB formation.

4.3 Models for fork-collapse

In the context of our experimental conditions, treatment of mouse embryonic fibroblasts ($ATR^{+/-}$) with a low dose of aphidicolin leads to polymerase slowing and greater tracts of ssDNA at replication forks. This can enhance formation of secondary structure by simple repeats, such as CAGAGG, on the exposed ssDNA. While present on either the leading or lagging strand of a replication fork, the structure on one strand exposes ssDNA on the complementary strand, which recruits RPA. This structure, particularly in the absence of the stabilizing function of ATR, can become susceptible to structure-specific nucleases mentioned above that catalyze cleavage and DSB formation. A scenario may arise whereby an intact structure on the cleaved strand could preclude efficient strand invasion and homologous recombination repair, leading to a persistent break. This model is described in Figure 36.

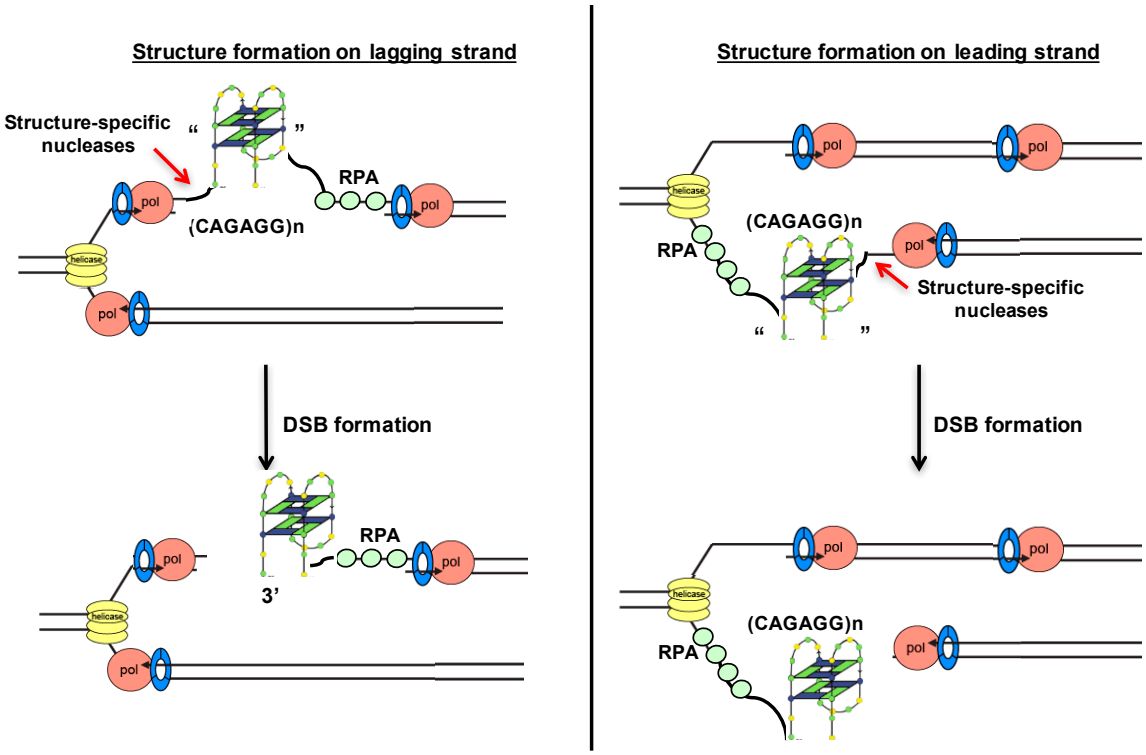

Figure 36. Model for fork collapse at $(CAGAGG)_n$ repeats. Putative structure formed by $(CAGAGG)_n$ repeats is shown, its presence depicted as either on the leading or lagging strand of the replication fork. The MCM2-7 helicase is depicted ahead of polymerases δ and ε of the lagging and leading strand, respectively. Red arrows indicate sites of cleavage by structure-specific nucleases after structure formation.

For those breaks that do not accumulate significant levels of RPA, exposure of ssDNA at replication forks by aphidicolin treatment can lead to extrusion of hairpins at sequences that contain inverted repeats, which can stall the polymerase. The helicase may continue to unwind DNA at the fork, leading to generation of hairpins on the complementary strand and complete stalling of the fork. Extruded hairpins or cruciforms can become susceptible to structure-specific nucleases that cleave the structure, leading to DSB formation. The resultant hairpin or cruciform structure can preclude the formation of RPA molecules at the collapsed fork due to limited exposure of ssDNA, thus accounting for insignificant RPA accumulation at these sites. This model is described in Figure 37.

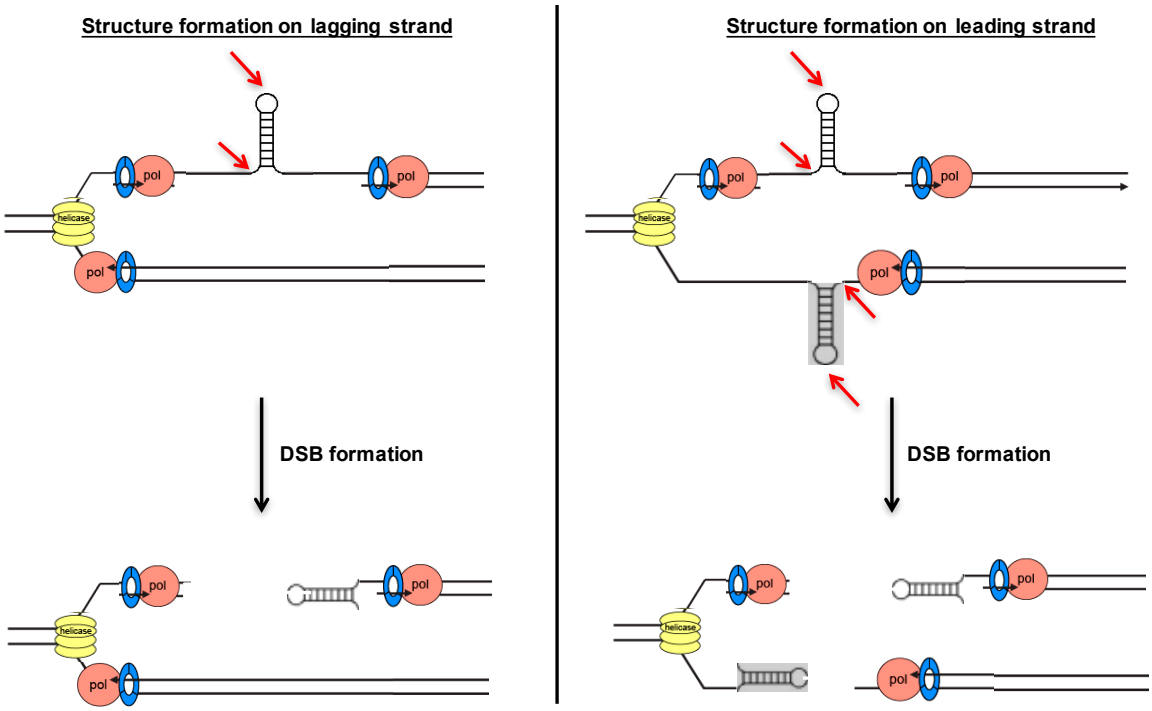

Figure 37. Model for fork collapse at inverted retroelements. Putative hairpin structures formed by inverted repeats is shown, its presence depicted as either on the leading or lagging strand of the replication fork. The MCM2-7 helicase is depicted ahead of polymerases δ and ε of the lagging and leading strand, respectively. Red arrows indicate sites of cleavage by structure-specific nucleases after structure formation.

4.4 Perspectives

It has been reported that activation of different oncogenes, such as Ras or cyclin E, leads to considerable genomic instability in the form of chromosome fragmentation that promotes a defined landscape of fragile sites in the genome (Miron et al., 2015). Interestingly, the landscape of fragile regions differs from that conferred by aphidicolin treatment and additionally varies between different activating oncogenes (Miron et al., 2015). Thus, diverse forms of replication stress are proposed to occur from distinctive circumstances of abnormal replication, affecting the selection and rate of fragile site expression (Miron et al., 2015). This suggests deviating effects on the genome under

unique conditions of replication stress. It would therefore be of great interest to identify the various landscapes of break sites promoted under different sources of replication stress, including those caused by activating oncogenes.

A variety of recent studies have pointed to “replication stress” as a cause of age-related pathologies; however, the mechanisms underlying these associations have remained obscure (Burhans and Weinberger, 2007). In one recent study, ribosomal repeats were implicated as a cause of stem cell aging (Flach et al., 2014). Other studies have implicated extreme replicative demand as being associated with age-related pathologies, such as tissue regeneration after injury (Burhans and Weinberger, 2007; Ruzankina et al., 2007; Ruzankina et al., 2009). It is interesting to speculate that repeat regions identified in these studies might cause replication fork collapse under urgent compensatory renewal and that these breaks and the ensuing DNA damage response could erode stem cell potential. Consistent with this possibility, triplet repeat expansions have been associated with pathologies related to those observed with advanced age (Orr and Zoghbi, 2007). Whether expansions of the simple tandem repeats described here will accelerate tissue degeneration with age and other age-related pathologies, such as cancer, has yet to be examined.

To conclude, our findings on highly specific localization of break sites generated by ATRi treatment has implications for cancer treatment. ATRi has entered Phase II clinical trials for the treatment of a variety of cancers. Although it is evident that the role of ATR in replication fork stability and cell cycle checkpoint control are central to the mechanism of ATRi action, difficult-to-replicate sequences that create a reliance on ATR function are part of that mechanism. The identification of such sites in different cancer types following single-agent treatment or combination with other targeted therapies and

chemotherapeutics may lead to new genomic pharmacodynamic biomarkers as well as a better understanding of the mechanism of action. In this light, it is intriguing to speculate that cancer cell-associated expansions of repetitive sequences that rely on ATR for stability may serve as predictive biomarkers of benefit from ATRi-based therapies. Accordingly, synthetic lethality with ATRi would not be based on defects in gene products, but rather with the genome itself.

References

Burhans BC, Weinberger M. (2007). DNA replication stress, genome instability and aging. Nucleic Acids Res. *35*, 7545-7556.

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, Zara F, Cañizares J, Koutnikova H, Bidichandani SI, Gellera C, Brice A, Trouillas P, De Michele G, Filla A, De Frutos R, Palau F, Patel PI, Di Donato S, Mandel JL, Cocozza S, Koenig M, Pandolfo M. (271). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science. *271*, 1423-1427.

Cimprich KA, Cortez D. (2008). ATR: an essential regulator of genome integrity. Nat. Rev. Mol. Cell Biol., *9*, 616–627.

Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, Lickley LA, Rawlinson E, Sun P, Narod SA. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. *13*, 4429-4434.

Dungrawala H, Rose KL, Bhat KP, Mohni KN, Glick GG, Couch FB, Cortez D. (2015). The replication checkpoint prevents two types of fork collapse without regulation replisome stability. Mol. Cell*. 59*, 998-1010.

Flach J, Bakker ST, Mohrin M, Conroy PC, Pietras EM, Reynaud D, Alvarez S, Diolaiti ME, Ugarte F, Forsberg EC, Le Beau MM, Stohr BA, Méndez J, Morrison CG, Passegué E. (2014). Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. Nature. *512*, 198-202.

Fu YH, Kuhl DP, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkert AJ, Holden JJ, Fenwick RG, Warren ST, Oostra BA, Nelson DL, Caskey CT. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell. *67*, 1047-1058.

Lahiri M, Gustafson TL, Majors ER, Freudenreich CH. (2004). Expanded CAG repeats activate the DNA damage checkpoint pathway. Mol. Cell*. 15*, 287-293.

Mandel JL, Heitz D. (1992). Molecular genetics of the fragile-X syndrome: a novel type of unstable mutation. Curr. Opin. Genet. Dev. *2*, 422-430.

McNees CJ, Tejera AM, Martinez P, Murga M, Mulero F, Fernandez-Capetillo O, Blasco M. (2010). ATR suppresses telomere fragility and recombination but is dispensable for elongation of short telomeres by telomerase. J. Cell Biol. *188*, 639-652.

Miron K, Golan-Lev T, Dvir R, Ben-David E, Kerem B. (2015). Oncogenes create a unique landscape of fragile sites. Nat Comuun. *6*, 7094.

Orr HT, Zoghbi HY. (2007). Trinucleotide repeat disorders. Annu Rev Neurosci. *30*, 575-621.

Pepe A, West S. (2014). MUS81-EME2 promotes replication fork restart. Cell Reports. *7*, 1048-1055.

Ragland RL, Patel S, Rivard RS, Smith K, Peters AA, Bielinsky AK, Brown EJ. (2013). RNF4 and PLK1 are required for replication fork collapse in ATR-deficient cells. Genes Dev. *27*, 2259-2273.

Ruzankina Y, Pinzon-Guzman C, Asare A, Ong T, Pontano L, Cotsarelis G, Zediak VP, Velez M, Bhandoola A, Brown EJ. (2007). Deletion of the developmentally essential gene ATR in adult mice leads to age-related phenotypes and stem cell loss. Cell Stem Cell. *1*, 113-126.

Ruzankina Y, Schoppy DW, Asare A, Clarck CE, Vonderheide RH, Brown EJ. (2009). Tissue regenerative delays and synthetic lethality in adult mice after combined deletion of Atr and Trp53. Nat Genet. *41*, 1144-1149.

Sarbajna S, Davies D, West S. (2014). Roles of SLX1-SLX4, MUS81-EME1, and GEN1 in avoiding genome instability and mitotic catastrophe. Genes Dev. *28*, 1124-1136.

Szakal B, Branzei D. (2013). Premature Cdk1/Cdc5/Mus81 pathway activation induces aberrant replication and deleterious crossover. EMBO J. *32*, 1155-1167.

Takai H, Smogorzewska A, de Lange T. (2003). DNA damage foci at dysfunctional telomeres. Curr. Biol. *13*, 1549-1556.

Walsh E, Wang X, Lee MY, Eckert KA. (2013). Mechanism of replicative DNA polymerase delta pausing and a potential role for DNA polymerase kappa in common fragile site replication. J Mol Biol. *425*, 232-243.

APPENDIX

Cell lines - MEF 4-3 cells (Smith et al., 2009); Tim KD MEF 4-3 cells (Smith et al., 2009); I-PpoI MEF 4-3 cells were generated by transducing retrovirus expressing the fusion I-PpoI restriction enzyme from the pBabe-ddIPpoI plasmid (Addgene plasmid #49052) into MEF 4-3 cells.

Cell treatments - MEF 4-3 cells were treated with either DMSO or 1 µM ATR-45 (Charrier et al., 2011) and 0.2 µM aphidicolin (Calbiochem, CAS 38966-21-1) for 18 hrs; 1µM ATR-45 and 0.2 µM aphidicolin for 9 hrs; 1 µM ATR-45 for 18 hrs; and 0.2 µM aphidicolin for 18 hrs. I-PpoI MEF 4-3 cells were treated with 1 µM Shield-1 (Wandless lab, Stanford), and 0.5 µM 4-hydroxytamoxifen (4-OHT, EMD) for 14 hrs to induce nuclear expression of I-PpoI. Parental MEF 4-3 cells were similarly treated with 1 µM Shield-1 and 0.5 µM 4-OHT for 14 hrs as a control.

Cell culture – All cells were grown in Dulbecco modified Eagle's medium (DMEM, Mediatech) supplemented with 10% fetal bovine serum (FBS, Benchmark, Gemini BioProducts), L-glutamine (2 mM, Mediatech), and streptomycin/penicillin (100 U/ml, Thermo Fisher Scientific).

RPA ChIP-Seq - For each immunoprecipitation reaction, 15 x 10^6 cells were trypsinized, collected, spun down and re-suspended in 25 mL PBS. Cell were fixed in 1% formaldehyde for 10 minutes at 37˚C and the reaction was stopped by adding glycine to 1% final concentration. The cell pellet was washed in 10 ml PBS and subsequently re-suspended in 1 ml cold PBS. The pellet was then lysed in lysis buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NPZ40, 0.25% Triton X-100) for 10 minutes on ice. The nuclei were recovered by spinning and washing twice (10 mM Tris-Cl

pH 8.1, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0). The nuclei were re-suspended in 1 ml shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris pH 7.6), and chromatin sheared using Covaris S220 to <4 kb using parameters according to the company hand book. Buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris pH 7.6, 11% Triton X-100, 1.1% Na-DOC) was added to 1/10 volume to keep the sample in Radioimmunoprecipitation assay (RIPA) buffer.

Dynabeads Protein A beads were pre-bound the night before by mixing 1 ml PBS, 10 µl 100 mg/ml BSA in PBS, 20 µg anti-RPA32 antibody (NA19L, Oncogene), and 10 µg bridging antibody, rotating overnight at 4˚C. The next day, the beads were washed as follows: 2x with 1 ml of RIPA buffer, 2x with 1 ml of RIPA buffer + 0.3 M NaCl, 2x with 1 ml of LiCl buffer (0.25 M LiCl, 0.5% NP-40, 0.5% NaDOC, store at 4˚C), 1x with 1 ml of TE (pH 8.0) + 0.2 % Triton X-100, 1x with 1 ml of TE (pH 8.0). The beads were then incubated with Proteinase K at 65˚C to reverse cross-link. DNA was extracted using phenol-chloroform and precipitated with ethanol/sodium acetate. Pellets were re-suspended in TE (pH 8.0) and processed for downstream qPCR analysis and NGS library preparation.

BrITL - For each BrITL reaction, ~2 x 10^6 cells were trypsinized and collected in an Eppendorf tube. Cells were washed with PBS, permeabilized in 0.1% Triton-X-100 in PBS for 5 minutes on ice and subsequently washed with 0.01% Triton-X-100 in PBS. Cells were incubated in a reaction containing 20 µM ddNTPs (Affymetrix, 77126) in 1X NEBuffer 2 for 5 minutes at 37°C. The reaction was stopped with 20 mM EDTA. Cells were washed four times with 0.01% Triton-X-100 in PBS before resuspending the cell pellet in a reaction mixture containing 2.5 mM $CoCl_2$ (Roche, 11243306001) and 27 µM biotin-16-ddUTP (Enzo Life Sciences, ENZ-42813) in 1X TdT buffer (Roche, 11243276001). Upon addition

of 150 units of TdT (Roche, 03333566001), the end-labeling reaction proceeded for 1 hour at 37°C. Cells were then washed twice with 50 mM EDTA in 0.01% Triton-X-100 in PBS.

To lyse the cell pellet, TNE buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 0.1 µM EDTA) was added together with 10% SDS and 10 mg/ml Proteinase K for incubation overnight at 37°C. The next day, genomic DNA was extracted using phenol/chloroform followed by ethanol/sodium acetate precipitation. The pellet was re-suspended in TE (pH 8.0). Sonication occurred in the Biorupter (Diagenode) for 2 minutes on medium setting to obtain 0.2-2 kb fragments.

After sonication, the samples were purified with Ampure XP beads (Beckman Coulter, A63880), utilizing 0.8x SPRI:DNA ratio. Washed and dried beads were incubated with EB buffer and left at room temperature for up to an hour before placing at 4°C overnight. The next day, the eluate was retrieved from the beads and brought up to 100 µl volume with TE. From this volume, 15 µl was aliquoted into a separate tube containing 85 µl TE and stored at 4°C to serve as the input. The rest of the sample was brought up to 200 µl with TE and proceeded to the next steps for retrieval.

Selection of biotin-labeled DNA fragments was performed with the Dynabeads KilobaseBinder kit (Life Technologies, 601-01). For this, 25 µl of streptavidin-coated magnetic beads were washed twice with 200 µl Binding buffer containing 5 µg/ml tRNA. The beads were then mixed in 200 µl sample plus 200 µl Binding buffer and left at room temperature on a rotating wheel for 2 hrs. The samples were then placed against a magnetic stand and the supernatant discarded. The beads were washed twice with wash buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl) with 5 minute rotations at room temperature for each wash. The beads were then transferred to a new tube containing

wash buffer with 4 µg/ml tRNA and subsequently washed in distilled autoclaved water twice. The washed and dried beads were then re-suspended in TE. To these samples, 20 µg boiled RNase A was added. Samples were incubated at 37°C for 30 minutes to remove RNA contaminants.

Elution of biotinylated fragments bound to streptavidin-coated magnetic beads occurred by adding 1% SDS and 1 mg/ml Proteinase K to the samples and incubating at 55°C overnight. The next day, DNA was purified by sequential phenol, phenol/chloroform, and chloroform extraction before subsequent ethanol/sodium acetate precipitation. The DNA pellet was re-suspended in 50 µl of TE.

qRT-PCR analysis: Real-time PCR was performed on the Applied Biosystems Real-time PCR detection system. For each primer set, 2 µl of retrieved DNA isolated as described above and 2 µl of the respective input were used. Primer sequences are described below:

RPA Peak Genomic Location (mm9)	Distance relative to 'CAGAG G' repeat region	Forward Primer (5' -> 3')	Reverse Primer (5' -> 3')
Chr7: 35943530 – 35946581	20 kb 5'	GCAAGCCATGGAACTCATCTA	TGGCATGAAGACACCAAGAG
	1,260 bp 5'	CCCTGGATGTGGTTGTCATTA	GGAACAGAACTGCTCACAAATG
	220 bp 5'	ACACCCAGTGAACCAAAGTATAG	CCACCAGTCTTGACTTACTCAC
	230 bp 3'	GACAAGTGCCTGAGGGATAAA	ACTCAGGAGACAGAGGGTTAT
	530 bp 3'	AGAATCCTCGGGTGGAAATG	CAGTGGCTCCTGTTGTGTAT
	20 kb 3'	GCTGAGGCAGGGATAGATTTG	AGGCTTGAGAAGAGTTGAAGATAAG
Chr6: 87722438 – 87728725	20 kb 5'	CCGACTGACCTGTGCTATTT	CATGTGGTTGCTGGGATTTG

	1,450 bp 5'	GTGATTGCCACCAAACCTAATG	CTGGGAAAGGCAGAGAACTAC
	285 bp 5'	CCCAGCACTTGGTAAGTAGAG	AACTCTCGCAGCCGTTTAT
	970 bp 3'	GCCCAAGGCTCTACCATAAA	GCCTGGAACCTACCTTCATAC
	20 kb 3'	CACAGACACGCACACATAGA	TGCTAGGGCCTGAGAATAGA
RPA Peak Genomic Location (mm9)	**Distance relative to 'CAGG' repeat region**	**Forward Primer (5' -> 3')**	**Reverse Primer (5' -> 3')**
Chr9:121838100-121841700	20 kb 5'	GTGACTTTATCTCCACCGTACTC	CTACTGGAGTGAACCTGGAATG
	460 bp 5'	GCAGCTTGGCTCTGTATCTAT	GTGTGTGTGTTCAGAGGTCTAT
	165 bp 5'	CCTGTCTCTGACACAATGTACTC	CCCAATACACCGCTCCTTT
	230 bp 3'	GCCTCTAGTGATACATCTCCTCTA	GATAAGGCCATCTGCAGTGAT
	360 bp 3'	CAAAGTGCCTCTCGCCTAAA	GCCTGGCTTCACAAGGATAG
	20 kb 3'	CCTTTGCAGGAAGTAGGTCAA	AGAGATGCCAGGAGCAAATC
RPA Peak Genomic Location (mm9)	**Distance relative to 'CAAAA' repeat region**	**Forward Primer (5' -> 3')**	**Reverse Primer (5' -> 3')**
Chr11:5641800-5644000	20 kb 5'	TCCTTCCATTTCTCTCCATTCC	CACGGTCAATTCGGACTTCT
	617 bp 5'	TGGGCTCTGCCTACTTACTA	GCCCTCAGAGAAGAAGAGAATG
	300 bp 5'	GGCTGGATTCAAGGCAGTAA	CCTCCCAAGTGCTGAGATTAAA
	440 bp 3'	TGGGCAGCGAAGATGAAAT	GCCAAGCCACTCCCTTATT
	20 kb 3'	GTGCAGAGACTATGGAGGAATG	GCAAAGGATCCTGGGTTGTA
RPA Peak Genomic Location (mm9)	**Distance relative to 'CACAG' repeat region**	**Forward Primer (5' -> 3')**	**Reverse Primer (5' -> 3')**

Chr17: 13737678-13751955	20 kb 5’	CTTTCTGTGTCCGTGTCCTATC	CACCAACCGTAGAATGGGTATC
	1,085 bp 5’	TTCAAGGTGAGAGAGATGGATTG	GCCTTCTCCTGTGCTGTATT
	600 bp 5’	GAAGTATCAGGCAGCAAGGAA	CTGACTTTGACGGCAGGATAA
	0 bp 3’	GCTGTCTCTGCTGTCTGTAATG	GCTACTCTTGTCCCATGTTTCC
	500 bp 3’	GCCTCCAGGTTGCCTAAATA	GCTCTCTGTTGCTATGGTGAA
	1,400 bp 3’	CCCGCTAATTCCTCTGAAGTC	TCTTCCCTGACTGGTCCTATT
I-PpoI site in rDNA sequence	**Distance relative to I-PpoI site**	**Forward Primer (5' -> 3')**	**Reverse Primer (5' -> 3')**
	11.5 kb 3'	CACGCTGTCCTTTCCCTATT	GACAGACCCAAGCCAGTAAA
	7 kb 3'	CTGAGAAACGGCTACCACATC	GCCTCGAAAGAGTCCTGTATTG
	780 bp 3'	ACAGCCTCTGGCATGTTG	GCCAATCCTTATCCCGAAGTTA
	70 bp 3'	CTAGCAGCCGACTTAGAACTG	CAGAAATCACATCGCGTCAAC
	60 bp 5'	CCTACCTACTATCCAGCGAAAC	CTACACCTCTCATGTCTCTTCAC
	120 bp 5'	GGGAAAGAAGACCCTGTTGAG	GGCCTCCCACTTATTCTACAC
	750 bp 5'	GAACGTGAGCTGGGTTTAGA	CTCTCGTACTGAGCAGGATTAC
	20 kb 5'	GAAACCAAAGCGACCTGAAAC	CAGCCATCTTGTCTGCTAACT

Bioinformatics -

Peak-calling: ChIP libraries were sequenced through Illumina HiSeq, generating 100 bp single-end sequencing reads. Adapter contamination in reads were trimmed using trimmomatic (Bolger et al., 2014) and reads were checked for quality control using fastqc (Leggett et al., 2013). Alignment was made to the mm10 reference genome using STAR Aligner with at most 3 mismatches (Dobin et al., 2013). Reads were initially allowed to be placed in up to 100 different genomic regions in order to later measure differences in

regional read accumulation between multi-mapping of a single read with up to 100 placements and mapping of a single read to its most likely genomic placement. In the context of these experimental regions, measuring the difference between tracks with reads that can potentially have up to 100 different placements and those with reads that are placed in their most likely home will reveal enrichment bias toward low complexity regions (i.e. if it is solely due to the low complexity nature of these regional sequences and not due to the experimental enrichment).

Reads were then filtered by mapq score 10 to keep the high-confident read mappings. De-duplication of reads in the aligned tracks took place to increase the complexity of the read population. Additional alignment-specific quality-control metrics were conducted, including strand-cross-correlation (Landt et al., 2012), finger-plots (Ramírez et al., 2016) to gauge mutual back-level of enrichment across samples, Pearson and Spearman correlations of genomic and enriched regions across samples ($\geq$0.6), principal component analysis (PCA) for clustering assessment, and a non-arbitrary estimate of ChIP signal over input tracks using an NCIS-generated normalization ratio (Liang and Keles, 2012). Black-listed regions in the mm10 genome were filtered out prior to peak-calling.

For enrichment analysis, the biological replicates and inputs of each experimental condition underwent an irreproducibility rate (IDR) analysis (Landt et al., 2012) from the ENCODE project with the MACS2 peak-calling program (Zhang et al., 2008) to give the final peak list per condition. IDR thresholds of >0.05 was used for self-consistency and comparison of biological replicates, and >0.005 for pooled-consistency analysis. Peaks that passed IDR thresholds were further filtered to select those with p-value $<10^{-3}$ and that were above 4-fold enriched over input. Regions within 2 kb of one another were merged.

The final peak list per condition was generated as a set intersection with and subtraction from the DMSO-control peak list.

Enrichment of complex repeats in RPA ChIP samples: Trimmed fastq reads from each RPA ChIP-Seq sample that overlapped with different families of complex repeats (LSU_rRNA, SSU_rRNA, tRNA, etc.) were counted for each family of repeats. These numbers were then divided by the total number of reads with at least one reported alignment in each sample. Values from different biological replicates in each condition (ATRi+aph^{18hrs} and UT) were averaged and normalized by the values calculated in the respective input samples. The resulting fold over input values for each family of complex repeats were graphed for each condition.

REQer: To understand simple repeat sequences that may be enriched in the experimental conditions relative to input, an assessment of the sequence presence within individual reads was performed. Reads in fastq files were labeled according to how many times a sequence occurs as a single unit (monomer), or as different tandem units, using a python script that incorporated regular expressions. This program was called REQer.

Contiguous simple repeat analysis of trimmed and de-duplicated RPA ChIP-Seq reads from combined biological replicates of each condition (ATRi+aph^{18hrs} and UT) was conducted by counting the total occurrences of a specified number of tandem monomers of each repeat within all reads. The frequency of the occurrences was measured as a percentage of total monomer count of the repeat present in the reads of the combined replicates. At each specified number of tandem monomers, the ratio of the ATRi+aph^{18hrs} value over its input and of the UT value over its input was calculated to obtain fold over input enrichment.

Non-contiguous simple repeat analysis of trimmed and de-duplicated RPA ChIP-Seq reads from combined biological replicates of each condition (ATRi+aph^{18hrs} and UT) was conducted by categorizing the reads by the total monomer count of each repeat within a read. Their frequency was calculated as the fraction of reads within the total number of reads from the combined replicates that contained the specified amounts of repeat monomers. At each specified number of total monomers per read, the ratio of the ATRi+aph^{18hrs} value over its input and of the UT value over its input was calculated to obtain fold over input enrichment.

P-values were obtained at 95% confidence interval using the Kolmogorov-Smirnov test between the distributions of ATRi+aph^{18hrs} and its input.

Fork-pausing –

Plasmid constructions: 630 bp of CAGAGG tandem repeats were cloned into the BspEI and BssHII site in the pML113 plasmid in opposite orientations for the origin-proximal insertion, and into the BamHI site for the origin-distal insertion. Randomized controls of the same nucleotide composition and length were similarly constructed.

In vitro assay: Templates for polymerase reactions were created by cloning $[CAGAGG]_{15}$ repeats into the MCS/BamH1 site of the pGEM3Zf(-) vector. Inserts in two orientations were isolated in order to purify ssDNA templates of both strands. As controls, randomized sequences of the same nucleotide composition and length were similarly cloned. Subsequently, a double G to T mutation at the 5' BamHI site (GGATCC) and a C to A mutation at the 3' BamHI site (CCTAGG) flanking the repeat insert were introduced, in order to disrupt the potential for G-quadruplex formation between the vector and insert sequences. For each construct, single-stranded DNA was isolated after R408 helper

phage infection of plasmid-bearing, SURE cells (*e14-(McrA-), Δ(mcrCB-hsdSMR-mrr)171, endA1, gyrA96, thi-1, supE44, relA1, lac, recB, recJ, sbcC, umuC::Tn5 (Kanr), uvrC [F' proAB lacIqZΔM15 Tn10]*; Agilent Technologies). Small ssDNA preparations from independent clones were sequenced to verify integrity of the insert prior to large scale purification of ssDNA templates. Repeat lengths longer than 15 units precluded our ability to rescue ssDNA of the correct sequence and/or length. DNA synthesis templates were created by hybridization of a ^{32}P end-labeled oligonucleotide (1:1 molar ratio) that initiates synthesis 68 nucleotides upstream of the repeat inserts. Reactions contained 100 fmol of primed ssDNA substrate, 400 fmol PCNA, 1700 fmol RFC, 20 mM Tris HCl, pH 7.5, 8 mM $MgCl_2$, 5 mM DTT, 40 µg/ml BSA, 150 mM KCl, 5% glycerol, 0.5 mM ATP, and 250 µM dNTPS, and were preincubated at 37°C for 3 min. Synthesis was initiated upon addition of 100 fmol purified 4-subunit recombinant human Pol δ4 (Zhou et al., 2012). Aliquots were removed at 3, 7, and 15 minutes, and reaction products were separated on an 8% denaturing polyacrylamide gel and quantitated using a Molecular Dynamics Phosphoimager. A control for the percent of primers productively hybridized to each primer/template substrate (% Hyb) was performed using excess Exo$^-$ Klenow polymerase, and a background control for primer impurities (no Pol) was performed by incubating unextended primer/template substrate in reaction buffer without the addition of polymerase. Dideoxy sequencing reactions were carried out simultaneously with the Pol δHE reactions, using the same primer/template substrates and Sequencase 2.0 (Thermo-Fisher). Total percent extension was calculated as the amount of total extended DNA molecules (corrected for percent hybridization and background) divided by this number plus the amount of corrected primer molecules. The number of DNA molecules within four regions were determined from the 15 minute reaction using ImageQuant software

quantitation: (R1) 68-11 bases 5' to the insert; (R2) 10-1 bases 5’ to the insert; (R3) the insert; and (R4) all bases 3’ to the insert up to and including the well. After background correction, the termination probability within each region was calculated as the [number of molecules within the region ÷ by the number of molecules within the region plus all longer molecules]. To normalize for the different sizes of Regions 1-3, each region’s termination probability was divided by the number of nucleotides under consideration. For example, the termination probability/nt for Region 1 = [molecules in R1 ÷ molecules in R1+R2+R3+R4] ÷ 58 nucleotides.

Ex vivo assay: The SV40-derived pML113, 114 and 115 vectors (Follonier et al., 2013) were gifts from Massimo Lopes (University of Zurich). For the ori-proximal vectors, a 630 bp fragment encoding CAGAGG tandem repeats was cloned into the pML114 and 115 plasmids using the MCS/BspEI and BssHII sites, creating plasmids with the repeats in two orientations. As controls, randomized sequences of the same nucleotide composition and length were created and similarly subcloned into pML114/115 vectors. For the ori-distal vectors, the repeats were cloned into the BamHI site of pML113, in two orientations. Subconfluent U2OS cells (ATCC) were transfected with 5µg vector DNA. To induce replication stress, cells were treated with 0.6 µM Aphidicolin, 24 hours post-transfection. For all experiments, DNA was isolated and replication intermediates purified 48 hours post-transfection, as previously described (Chandok et al., 2011). Purified DNAs were digested with DpnI, EcoRI, and EcoN1 (ori-proximal) or DpnI, PpuMI, and SacII (ori-distal) restriction enzymes. Replication intermediates were separated by neutral/neutral electrophoresis DNA using an 0.4% TBE agarose gel (1V/cm, 14 hr, room temperature) in the first dimension and a 1% TBE agarose gel (4 V/cm, 6-8 hr, 4˚C) in the second dimension. (Friedman and Brewer, 1995). Gels were transferred to Hybond-N+

membranes (Amersham) and hybridized with the indicated ^{32}P-labeled DNA probe, according to manufacturer's instructions. For quantitation, blots were scanned using a Phosphoimager, and following background correction, the Replication Fork Barrier (RFB) index was calculated for each 2D image as: RFB = [(double Y arc spike) ÷ (simple Y descending arm)].

Biophysical Characterization of DNA Secondary Structure -

DNA and buffers for structural studies: All DNA samples were ordered from IDT (Texas, USA) as HPLC purified samples, dissolved in water at 1 mM final concentration and stored at -80°C. Samples were diluted to the desired concentration into final 10 mM lithium cacodylate buffer pH 7.2 supplemented with 100 mM KCl and 2 mM $MgCl_2$ (100K2Mg buffer). Samples were annealed at 90°C for 5 minutes, cooled slowly to room temperature over the course of 3-4 hrs and equilibrated overnight at 4°C. All samples were examined with circular dichroism (CD) for consistency in folding.

Circular Dichroism (CD) wavelength scans: All experiments were performed on an AVIV 410, AVIV 435, or a Jasco 815 spectropolarimeter with a Peltier heating unit using 1 cm quartz cuvettes. The accuracy of the external temperature probe was ± 0.3 K. Each CD trace was an average of 3 - 5 scans collected from 220 to 330 nm with 1-2 nm bandwidth, 1 nm step, 1 second averaging time at 4ºC. CD data were converted to molar ellipticity according to the following formula: $\Delta\varepsilon = \frac{\theta}{3.298\times10^4\times l\times C}$, where θ represents CD signal in mdeg, *l* is cuvette pathlength in cm; and C is DNA concentration in M (per strand). Zero correction was performed using the average of the data from 320-330 nm. When

necessary, the resulting curves were smoothed using a Savitzky-Golay filter with a 13-point quadratic function.

CD melting: Thermal denaturation experiments were collected at 261 nm (or max in the CD wavelength scan) with 2 nm bandwidth, 5 second equilibration time, 1°C step, and 15 or 20 second averaging time, and 10 nm bandwidth. The samples were heated from 4 to 95°C, maintained at 95°C for less than 5 minutes, and then the temperature was decreased to 4ºC at the same rate. The cooling step was included to determine the reversibility of folding/unfolding process. Superposition of melting and cooling data suggested that the folding process was reversible. Thus, melting data were analyzed assuming a two-state model with constant ΔH (Ramsay and Eftink, 1994). This model suggests that at any point during melting or cooling only folded and unfolded DNA is present (no intermediates). Starting and final baselines (assuming to be linear), melting temperature and enthalpy of unfolding were adjusted to get the best fits.

Molecularity of DNA structures via UV-vis melting: Studies were performed on a Secomam Uvikon XL spectrophotometer thermostated with an external Julabo F12-ED waterbath. Samples were prepared in 100K2Mg buffer with concentrations ranging from 1.4 to 23.6 μM for CA5 and from 0.3 to 7.6 μM for CA10; targeted concentration ratio between the most dilute and the most concentrated sample was ~20. CD scans were collected before melting to assure correct DNA folding. Samples were placed in cuvettes with 1.0 or 0.2 cm pathlength depending on strand concentration. The temperature was measured with the temperature sensor inserted in the cuvette holder right next to DNA samples. Cuvettes were equilibrated at 4°C for at least 20 minutes, then the temperature was increased to 80°C with 0.2°C/min temperature ramp and 1 second averaging time. Subsequently

samples were cooled back to 4°C using the same parameter. Data were collected at 260, 295, and 335 nm (the latter wavelength was used as a reference to factor out instrument fluctuations). The data were corrected for absorbance at 335 nm, and were either fit using a two-state model with constant ΔH (same as for CD data) or analyzed using a derivative method (yields Tm, but no thermodynamic parameters). In the latter case, minima or maxima on the first derivatives of the melting curves were used as Tm.

Polyacrylamide gel electrophoresis: Native PAGE gels were typically prepared at 12% polyacrylamide in 1×TAC (50 mM Tris Acetate pH 7.3) buffer supplemented with 3 mM $MgCl_2$. Running buffer consisted of 1×TAC with 3 mM $MgCl_2$. Gels were cooled with a water bath and premigrated for 30 minutes at 140 V. Each sample of 10 μL contained ~3 μg of annealed oligo in 100K2Mg buffer to which 3 μL of 50% w/v sucrose was added immediately prior to loading. Oligothymidylate markers 5′ dT_n (where n = 15, 24, 30, 57 or 60, and 90) as well as a 76-nt tRNA were used as internal migration standards. Typically, gels were run for 3-4 hrs at 140-300V; gel temperature did not exceed 16 °C. Gel was stained using Stains-All and de-colored under visible light. Gels were visualized on ETNA-NS ChemiBis 3.2 gel visualization device (using lower light, 580 nm filter) or with iPhone 5 camera.

References:

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. *30*, 2114–2120.

Chandok GS, Kapoor KK, Brick RM, Sidorova JM, Krasilnikova MM. (2011). A distinct first replication cycle of DNA introduced in mammalian cells. Nucleic Acids Res. *39*, 2103-2115.

Charrier JD, Durrant SJ, Golec JM, Kay DP, Knegtel RM, MacCormick S, Mortimore M, O'Donnell ME, Pinder JL, Reaper PM, Rutherford AP, Wang PS, Young SC, Pollard JR. (2011). Discovery of potent and selective inhibitors of ataxia telangiectasia mutated and Rad3 related (ATR) protein kinase as potential anticancer agents. J Med Chem. *54*, 2320-2330.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics. *29*, 15-21.

Friedman KL, Brewer BJ. (1995). Analysis of replication intermediates by two-dimensional agarose gel electrophoresis. Methods Enzymol. *262*, 613-627.

Follonier C, Oehler J, Herrador R, Lopes M. (2013). Friedreich's ataxia-associated GAA repeats induce replication-fork reversal and unusual molecular junctions. Nat Struct Mol Biol. *20*, 486-494.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. *22*, 1813-1831.

Leggett RM, Ramirez-Gonzalez RH, Clavijo BJ, Waite D, Davey RP. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front Genet. *4*, 288.

Liang K, Keles S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics. *13*, 199.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160-165.

Ramsay G, Eftink MR. (1994). A multidimensional spectrophotometer for monitoring thermal unfolding transitions of macromolecules. Biophys J. *66*, 516-523.

Ruzankina Y, Pinzon-Guzman C, Asare A, Ong T, Pontano L, Cotsarelis G, Zediak VP, Velez M, Bhandoola A, Brown EJ. (2007). Deletion of the developmentally essential gene ATR in adult mice leads to age-related phenotypes and stem cell loss. Cell Stem Cell. *1*, 113-126.

Smith KD, Fu Ma, Brown EJ. (2009). Tim-Tipin dysfunction creates an indispensable reliance on the ATR-Chk1 pathway for continued DNA synthesis. J Cell Biol. *187*, 15-23.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.